DOCUMENT RESUME

ED 068 916 CS 000 267

AUTHOR	Leibert, Robert E., Ed.
TITLE	Diagnostic Viewpoints in Reading.
INSTITUTION	International Reading Association, Newark, Del.
PUB DATE	71
NOTE	140p.
AVAILABLE FROM	International Reading Association, 6 Tyre Avenue, Newark, Del. 19711 ($4.00 non-member, $3.00 member)
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	*Conference Reports; Diagnostic Teaching; *Diagnostic Tests; Elementary School Students; Evaluation; Informal Reading Inventory; Measurement Instruments; Reading Comprehension; *Reading Diagnosis; Reading Difficulty; Reading Instruction; *Reading Research; *Reading Tests; Secondary School Students; Syntax

ABSTRACT

A collection of papers delivered during the Fifteenth Annual International Reading Association Convention is presented which represents a variety of views on diagnosis and/or on the manner in which diagnistic information is interpreted. The papers have been arranged into three sections. The first deals with the importance of diagnosis, presents some methods for collecting and interpreting data about reading progress, and describes a plan for bringing about changes in reading performance. The second section treats tests and testing and provides information on ways a teacher can use tests. Included are an analysis of several diagnostic tests currently available, a discussion of problems and solutions in utilizing both standardized and informal tests, and a description of the development of a diagnostic test. Section 3 is composed of four reports analyzing data to shed light on the relation between intelligence and reading improvement, the stability of reading achievement, and critical evaluations of methods for determining levels of achievement. The papers are arranged in a way that makes the monograph easy to use, especially the treatment of the statistical studies. Tables and references are included. (This document previously announced as ED 047 909.) (Author/DH)

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED 068916

DIAGNOSTIC VIEWPOINTS IN READING

ROBERT E. LEIBERT, *Editor*
University of Missouri at Kansas City

INTERNATIONAL READING ASSOCIATION
Newark, Delaware 19711

FILMED FROM BEST AVAILABLE COPY

CS 000 267

INTERNATIONAL READING ASSOCIATION
OFFICERS

1970-1971

President: DONALD L. CLELAND, University of Pittsburgh, Pittsburgh, Pennsylvania

President-Elect: THEODORE L. HARRIS, University of Puget Sound, Tacoma, Washington

Past President: HELEN HUUS, University of Missouri, Kansas City, Missouri

---

DIRECTORS

Term expiring Spring 1971

William K. Durr, Michigan State University, East Lansing, Michigan
Mildred H. Freeman, Urban Laboratory in Education, Atlanta, Georgia
Ethel M. King, University of Calgary, Calgary, Alberta

Term expiring Spring 1972

Thomas C. Barrett, University of Wisconsin, Madison, Wisconsin
Constance M. McCullough, San Francisco State College, San Francisco, California
Eileen E. Sargent, Nicolet Union High School, Milwaukee, Wisconsin

Term expiring Spring 1973

Marjorie S. Johnson, Temple University, Philadelphia, Pennsylvania
Robert Karlin, Queens College, City University of New York, Flushing, New York
Olive S. Niles, State Department of Education, Hartford, Connecticut

---

Executive Secretary-Treasurer: Ralph C. Staiger, University of Delaware, Newark, Delaware

Assistant Executive Secretary: Ronald W. Mitchell, International Reading Association, Newark, Delaware

Publications Coordinator: Faye R. Branca, International Reading Association, Newark, Delaware

Library of Congress Catalog Card Number: 73-142551
Copyright 1971 by the International Reading Association, Inc.

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
International Reading Association
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

Contents

Foreword v

Introduction vii

DIAGNOSTIC STRATEGIES

1 Reading Diagnosis—The Essential Ingredient in Teaching Every Child to Read

EVELYN JAN-TAUSCH develops the theme that diagnosis by the teacher is a necessary component of successful instruction. Examples of learner variables, the reading process, and factors affecting performance in the early stages of reading are given to illustrate the concept.

8 Strategies for Evaluating Reading Programs: Elementary Level

MARY AUSTIN describes the effect of more precise statements of objectives on the teaching task and on the outcomes of evaluation. Two systems for collecting and interpreting data are presented.

19 The Diagnostic Teaching of Reading

WANDA BREEDLOVE discusses the test-teach-test cycle as a viable strategy for diagnostic teaching. The emphasis is upon bringing about desirable changes in reading performance.

TESTS AND TESTING

31 How the Classroom Teacher Can Use a Knowledge of Tests and Measurements

MARVIN GLOCK provides an introduction to this section with his overview of two key concepts in measurement, those of validity and reliability. In addition, some of the factors affecting the interpretation of gains are discussed.

41 Strategies of Measuring Student Understanding of Written Materials

FRANK GUSZAK discusses problems and solutions connected with informal assessment of comprehension. The sample questions and tasks suggest instructional implications rather than simply determining the success or failure of the reader's ability to complete the task.

48 The Development of a Diagnostic Test of Synatactic Meaning Clues in Reading

ALBERT MARCUS provides a view of the development of a test. He also demonstrates how linguistic principles may contribute to the measurements of comprehension.

64 What Do Diagnostic Reading Tests Really Diagnose?

CAROL WINKLEY analyzes nine diagnostic tests in terms of the information each supplies to specific performance areas. Three tables provide a valuable summary of the major analyses made.

RESEARCH AND REFLECTION

81 Predicting True Reading Gains after Remedial Tutoring

ANITA DAHLKE reports on an analysis of data collected on clinic subjects to determine the relationship between "true reading gain" and subtests on the Wechsler Intelligence Scale for Children. The findings should encourage some experimental studies to help establish the causes of the key relationships identified.

103 A Longitudinal Study of Constancy of Reading Performance

KENNETH HOPKINS and GLENN BRACHT focus their investigation on the stability of reading achievement over an eleven grade span. The procedures and results have a bearing on the current move toward instructional accountability.

113 A Comparison of Formulas for Measuring Degree of Reading Disability

ALBERT HARRIS presents a case for a new reading expectancy quotient considered to be particularly useful in research, and with pupils who do not evidence clear-cut reading retardation.

121 The Validity of the Instructional Reading Level

WILLIAM POWELL reviews some of the recent criticisms of the criteria being used for determining instructional level with Informal Reading Inventories. Answers to three questions guide this inquiry into the criteria for and interpretation of instructional level.

Foreword

THE THEME of the International Reading Association's Fifteenth Annual Convention was *Reading and the Individual.* Attention was focused on the importance of adapting the methods and materials of reading to peculiar needs of the individual in order to help each person reach his highest potential.

One of the strands running throughout the convention was the emphasis on diagnosing reading problems so that the teaching can be synchronized with the deficiencies and thus achieve greater pupil improvement than when teaching is focused on general problems. This pinpointing of lacks in a pupil's reading skills forms the starting point for teaching that is truly diagnostic.

The papers in this volume have been arranged into three sections. The first deals with the importance of diagnosis, presents some methods for collecting and interpreting data about reading progress, and describes a plan for bringing about changes in reading performance.

The second section treats tests and testing and provides information on ways a teacher can use tests. Included are an analysis of several diagnostic tests currently available, a discussion of problems and solutions in utilizing both standardized and informal tests, and a description of the development of a diagnostic test.

Section three is composed of four reports analyzing data to shed light on the relation between intelligence and reading improvement, the stability of reading achievement, and critical evaluations of methods for determining levels of achievement.

Dr. Robert Leibert has arranged these papers in a way that makes the monograph easy to use, especially his treatment of the statistical studies. The International Reading Association is pleased to present this practical and readable volume.

However, the book would be of little value, regardless of its quality, unless those who read it put into practice the ideas and suggestions it contains. Only when classroom practice reflects the best in teaching will real progress be made. It is hoped that those

who teach reading will find this material helpful as they work to improve their own methods of teaching.

HELEN HUUS, *President*
International Reading Association
1969-1970

Introduction

DIAGNOSTIC VIEWPOINTS IN READING is a collection of papers delivered during the Fifteenth Annual IRA Convention at Anaheim. These papers present a variety of views on diagnosis and/or on the manner in which diagnostic information is interpreted.

Two ideas have been incorporated in arranging these papers to increase the readability of the volume. First, a statement or overview is included in the table of contents for each paper. Second, the two formal research papers in the third section have been separated into two parts. Part one includes the introduction, statement of the problem, methods and summary, and conclusion. Part two contains the description of the data analysis and related tables. The intent is to position information where it is most useful to the reader.

R.E.L.

The International Reading Association attempts, through its publications, to provide a forum for a wide spectrum of opinion on reading. This policy permits divergent viewpoints without assuming the endorsement of the Association.

DIAGNOSTIC STRATEGIES

Reading Diagnosis—The Essential Ingredient In Teaching Every Child to Read

EVELYN JAN-TAUSCH
Glen Ridge, New Jersey, Public Schools

THE PROBLEM of attempting to add anything significantly new or meaningful to the extensive literature already existing in the area of reading diagnosis is more than challenging—it is frustrating. The writer has worked in the area of reading instruction for her entire professional career and has been a conscientious reader of the professional literature that continues to pour forth in regular profusion. Very little of the material devoted to diagnosis seems either to dispute the findings of the first major researchers or to add to their dimensions. A good deal of it, frankly, strikes one as repetitive and concerned with going a very long way 'round to make some fairly obvious discoveries. Reports from the expert's diagnosis of a pupil's reading problem are too often confined to a list of the things he cannot do; e.g., "consistently omits word endings . . . confuses short *e* sound and short *u* . . . reads two years below grade level. . . . Too often the classroom teacher has already discovered these facts and greets the long-awaited diagnostic report with restrained enthusiasm. Why she hasn't been doing something about her own findings is, of course, a whole other subject in itself but one that does, the writer believes, touch at the very heart of our current educational problems. This personal conviction dictated the title "Reading Diagnosis—The Essential Ingredient in Teaching Every Child to Read" which expresses a philosophy that teachers have accepted verbally but have not generally incorporated into classroom practice.

The author believes that we have erred to some extent and continue to do so by making reading diagnosis so much of a *specialist* function that the technique stands in very real danger of becoming divorced from the classroom teacher's concept of her responsibility and of what the total act of teaching reading must and does include.

A specific example comes to mind that may help to illustrate this particular point. In her 1951 book *Growing into Reading,* Marion Monroe describes a language abilities evaluation technique that enables the teacher to chart each pupil's language pattern on one of five sequential levels in each of the following areas:

1.	Expressiveness	fluency and quantity
2.	Meaning	naming; description; inferential interpretation; narrative interpretation; evaluative interpretation
3.	Sentence Structure	isolated words; simple sentence; simple sentence with compound subject, predicate, or object; compound sentence with conjunction other than AND; complex sentence containing one dependent clause; complex sentence containing more than one dependent clause
4.	Word Meaning	cannot point to or define a pictured word upon request; can point to but not define verbally; defines by giving *use;* defines by description; defines by giving the generic class
5.	Qualities of Speech	voice tone; articulation; rhythm

Much professional interest has been directed in current years to the Illinois Test of Psycholinguistic Abilities (ITPA) and its value in reading diagnosis. In general, however, such testing has been considered the private domain of the psychologist, the speech therapist, or the reading clinician. The great advantage of the Monroe diagnostic tool is that it is assumed that it can and will be used by the *classroom teacher.* Familiarity with this strategy sharpens the teacher's awareness of what is involved in her own teaching objectives as well as what the specific language strengths and weaknesses are of *each* member of her class. It will be noted by those familiar with both tests that many of the same language areas are explored. The major difference is in *who* is going to use the instrument. This writer feels that this is a *difference* that makes a *difference* insofar as teaching children to read is concerned.

It is an interesting fact that while *diagnosis* continues to move in the direction of the specialist, *implementation* of the diagnostic prescription increasingly is held to be the job of the classroom teacher.

While one might choose to regard this pattern as perfectly analogous with the doctor's prescribing and nurse or parent's ladling out the medicine at the times and in the dosages indicated, there is a basic and highly significant contradiction. The *teacher* is an integral part of the *process* of teaching youngsters to read . . . the transmuting medium through which any prescription passes.

Most of the teaching of reading in our public schools is done by the classroom teacher. The lack of diagnostic teaching that prevails too often is due either to the teacher's not knowing how to diagnose or to the feeling that it is someone else's function to do so. In the instance of the first condition, it is possible for a trained observer to see an experienced first grade teacher miss the diagnostic implications in a child's reading behavior as follows: A pupil had been working for several sessions on word recognition in a remedial preprimer. Each time the pupil was asked in the lesson being observed to point to the word *brother,* she touched the illustration which pictured a young man. The teacher had previously stated that she could not understand why in the reading group the child, "learned a word one day and completely forgot it the next." The teacher evidenced no awareness that the basic reading concept of connecting a sound with its corresponding *alphabetic* representation had not yet been learned. It was this concept, not the word *brother* that had to become the instructional objective. The child was operating on another level, using a different set of visual symbols to cue her in to the meaning of the auditory sound.

Guy Bond's insightful statement (*1*) is applicable in too many classroom situations: "Many serious disabilities are simply the result of minor confusions which have been allowed to continue and pile up." Diagnosis at the *teaching* level is desperately needed if this situation is ever to be corrected. Diagnosis of Reading Difficulties has become over the years a course reserved for those who move into the specialized courses at the graduate level. (In some cases these specialists have never been "tainted" with actual classroom teaching.) The need for the elementary classroom teacher to be a diagnostician of reading problems remains, however, as a pressing and relevant problem.

Surely no one would contradict the rightness of another of

Bond's statements in the same article to the effect that "the effectiveness of diagnostic teaching is based upon the extent to which the teacher knows each child within the classroom . . . each child's capacities, his physiological condition, his emotional and social adjustments, his interests, attitudes and drives . . . his general level of reading ability. . . ." Tremendous as this task may seem (and it is a very large order) there is equal need for the teacher to know thoroughly what is involved in the reading process and to be able to determine the sequential order in which skills must develop. The twin requirements for reading diagnosis by the teacher are to know the learner *and* the reading process. One is tempted to add a third—what Goldhammer (*2*) has referred to as "an intelligent evaluation of his own teaching behavior."

In knowing the learner the teacher needs to analyze those factors which identify the learner in the learning situation:

1. At what stage of language development is the child? Is the child still having difficulty responding to vocal directions or is his difficulty localized in the vocal expression of his thoughts? Is his vocabulary very limited when he is compared with other children his age? Is his experiential background so limited that he has not had the need to find and use words to express his responses and reactions?
2. What models does he have at home for emulation? Are his parents and older siblings very limited in their uses of language to solve their personal problems and to provide recreation for themselves? Does the child identify with parents and older siblings whose life styles are characterized by out-of-doors activities, who need motor activities to feel comfortable, and who seek immediate gratification of their efforts rather than patiently await the accomplishment of long-term goals?
3. What motivates the child? Is he compulsive in his actions or is he a strategic thinker? What are his true interests? With whom and with what does he identify?
4. Are the child's sensory modalities functioning well? Does

he have adequate auditory discrimination? Is his peripheral hearing good? Does he have accurate visual discrimination? Does he respond appropriately to a combination of sensory stimulation concurrently received?

In analyzing the sequential steps of the reading process the teacher needs to understand that the following steps have a dependency relationship and a developmental structure:

Level 1. Awareness that speech sounds express thoughts and that thoughts can be expressed by speech sounds.

Level 2. Ability to manifest this awareness through appropriate action. (Anyone who can carry on a very simple conversation exhibits this awareness and ability.)

Level 3. Awareness that written symbols can describe sounds and, conversely, that sounds can be represented by letter symbols. (Pronouncing or pointing to the correct letter of the alphabet out-of-sequence and on demand would be indicative of this ability.)

Level 4. Awareness that written letter symbols and letter combinations can elicit thoughts and the ability to respond to such written letters symbols. (Correctly responding to signs such as stop, stand, sit, etc., is indicative of this ability.)

Level 5. Awareness that written letter and word-symbol *combinations* have a relationship which transcends the sound and meaning of *individual* letter and word symbols which comprise the combination. (This ability is manifested when the child even once correctly shifts the sound and meaning of words and word combinations to accord with context.)

Levels 4 and 5 can be more simply described by saying that they refer to the child's ability to bring sound and meaning *to* word symbols and the ability to derive sound and meaning *from* word combinations. An example is the child's ability to read correctly (orally with proper intonation; silently with correct comprehension) the following:

Lead is a heavy metal. He was asked to *lead* the march.
The *bank* was built on the *bank* of the river.

Teachers need to make the distinction between the basic reading act and its application—tasks which are involved in the reading process—when developing instructional objectives for teaching. The instructional objective has to be based, furthermore, upon accurate assessment of the individual pupil's needs. Is he still at Level 3 in terms of his need to develop the basic act of reading *or* is he unable to apply reading to other content areas? The school curriculum demands that the pupil apply the reading act to situations that use the basic skill as an expression of one's ability to *think* in the various subject areas. Ability to think is directly affected by one's intellectual capacity and one's experiential environment.

Teachers who have this kind of diagnostic insight can better understand and provide for that which the child needs to perform either in the achieving of the basic skill or in its application. When *providing* for it is an impossible task, then teaching plans must include compensation. The teacher may discover, for instance, that some of the factors which affect the quality of the pupil's ability to perform the basic reading act are not being afforded the attention they deserve, e.g.:

1. The child's development of the concept of self
2. The child's need to communicate through the use of language
3. The child's auditory vocabulary
4. The child's experiential background
5. The child's ability to cope with symbols

Similarily the child who poorly applies the reading act in meeting the demands of specific subject teachers may need *not* further instruction in learning to read but more pertinent learning experiences in the development of mathematical concepts or scientific reasoning or more relevancy in terms of his interests and experiences in the areas of literature, history, or economics. He may need programed instruction, multisensory approaches, concrete manipulative materials, or any one of many different instructional materials and

methods suited to his individual learning style. It will take diagnosis of the continuing and knowledgeable kind to enable the teacher to know the *when* and *what*.

There seems to this writer no better way, perhaps no other way, of reaching the goal of the '70s . . . every child a reader . . . than through equipping teachers to do a diagnostically-oriented job of instruction. It is necessary to put diagnosis back *into* the classroom. Reaching the moon has already proved an attainable objective, a fact proving once more that the universe within the mind of man is far more difficult to chart and navigate than the starry reaches of what's out there.

REFERENCES

1. Bond, Guy L. "Diagnostic Teaching in the Classroom," in Dorothy DeBoer (Ed.), *Reading Diagnosis and Evaluation*, 1968 Proceedings, Volume 13, Part 4. Newark, Delaware: International Reading Association, 1970.
2. Goldhammer, Robert. *Clinical Supervision*. New York: Holt, Rinehart and Winston, 1969.

Strategies for Evaluating Reading Programs: Elementary Level

MARY C. AUSTIN
University of Hawaii

TODAY, few educators fail to recognize that continuous, functional assessment is an essential ingredient for the total school program. This statement is true for a number of reasons, among them being: the post-Sputnik years which brought massive financial support for new courses in science and mathematics, the funding of Title 1 projects for children from impoverished homes, and the demands of influential citizens for appraisals of school offerings—all of which require "hard data" for the purpose of making educational improvements. Nor, for the most part, do educators fail to perceive that evaluation should encompass a broad spectrum of formal and informal procedures although, in practice, informal analysis with its reliance upon intuitive judgments and casual observations tends to take priority over such formal measures as standardized testing, checklists, and structured visitations. What school people often fail to realize, however, is that educational evaluation varies tremendously in quality from the highly discerning, accurate, and illuminating to the superficial, distorted, and limited, depending upon precisely stated objectives, carefully selected appraisal procedures, and the skill of those individuals who undertake the evaluation acts description and judgment.

Current Dissatisfactions

Dissatisfaction with some evaluative procedures is reasonable. Many achievement tests assess fact-finding abilities rather than the impact of instruction upon the acquisition of skills and understandings. Furthermore, achievement test results permit dissimilar interpretations. Behavioral data, admittedly desirable, are expensive to

gather and frequently do not answer the questions educators are asking. Checklists and questionnaires can be ambiguous. Evaluation teams can lack needed expertise in conducting interviews and visitations. And educators can avoid, through fear of public criticism, comprehensive evaluation plans which attempt to measure the complexity and importance of their programs.

One source of dissatisfaction is the testing of students' vocabularies. Administrators sometimes ignore the fallibility of tests as sole determinants of student abilities while classroom teachers, on the other hand, often become perplexed by problems of assessing vocabulary growth. The problem, moreover, may be complicated by the fact that many intelligence tests are essentially tests of vocabulary. As a case in point, the antonyms and analogies on such instruments as the Scholastic Aptitude do measure intelligence in that they test a student's ability to form abstract concepts and to reason from these concepts. But, as Hamlet said, there is a "rub." The brightest individual in the world cannot grasp an analogy if that individual is unfamiliar with the words used to illustrate the concept which is being tested by the analogy.

The complicated nature of language development makes its assessment difficult. Two major question arise: What vocabulary should be tested? What are "typical" vocabularies of children at given ages? Finding answers to these questions becomes important because many schools use the results of vocabulary-based intelligence tests to assign students to certain academic programs or tracks. Obviously, the findings of such tests need to be interpreted with caution. Makers of these instruments assume that all pupils share an equal exposure to the test items, and this assumption is patently untrue. Because of various educational, physical, and environmental facts and the part that these factors play in the vocabulary development of different individuals, "typical" vocabularies are next-to-impossible to ascertain.

Another problem embedded in the issue of testing pupil vocabularies arises from the deceptively simple question, "What is knowing a word?" If a child reads the word *rock* on a vocabulary test and thinks only *stone*, not *to sway*, he may miss the test item. But, is it fair to say that he does not know the word? In a very real

sense, no word has any meaning in isolation. A word has, of course, a certain inferential meaning; but the true meaning comes from its use in context, where the meaning is unique for that given set of circumstances.

Because words are merely symbols which people abstract as they use them, vocabulary knowledge may be said to exist on a continuum. Very few people know any word completely enough to be at the upper end of the continuum; hence, most words exist in their vocabularies in a "twilight zone" of partial meaning (*1*). It is virtually impossible, then, to determine at which point along the continuum any individual can be said to "know" a word.

A second source of dissatisfaction with evaluation procedures can be understood in relation to the number of interpretations which result from the same data. For example, the discovery that 75 percent of a group of inner-city children scores two years below grade placement on a vocabulary test may suggest to one teacher that she needs to turn to a number of word lists to develop ten new words a day until such time as the pupils are given an alternate form of the test. To another teacher, these results may suggest the limitations of general language development of ghetto children and the need to offer activities which extend and enrich experiences by a variety of direct and vicarious approaches. To an administrator who looks at the same data, the deficits of these pupils may lead him to apply for federal funds for a remedial reading programs. Meaningful, accurate interpretations of evaluative data can lead to improved reading programs, but many school systems have not yet reached this stage in their progress toward better assessment strategies.

Changes in Evaluative Practices

During the past twenty-five years, a number of innovations have occurred in evaluative practices. Changing concepts, different procedures, and new instruments of appraisal are direct results of a growing concern with behavioral psychology and the acceptance of a variety of learning theories. Because education is essentially a process of shaping or modifying behavior, evaluation must include

examination of behavior in its broadest sense with emphasis upon the interrelated aspects of thinking, feeling, and overt actions. The evolution of American society also plays a prominent role in changing evaluation techniques by creating new conditions for education which make possible different approaches in working with children. Such changes are inevitable because new educational practices require different evaluative procedures.

Neither can one debate the fact that marked changes in collecting and utilizing data have occurred, particularly during the 1960s. Giant electronic computers followed by newer minisized ones now collate and analyze information at incredible speeds. But technology notwithstanding, school people must still make decisions about what data to collect and how to use them. And then these data must be placed in proper perspective through relevant interpretation by the groups concerned.

Undoubtedly, one of the major changes during recent years centers around the concept of evaluation as one that involves more than the collection and analysis of information. Generally speaking, evaluation is a four step process: 1) stating purposes according to the needs of individuals, a community, and a changing society; 2) obtaining evidence as to how well these purposes are being realized; 3) interpreting the collected information; and 4) redefining goals, establishing new purposes, and planning appropriate programs to achieve the modified purposes.

Stating Purposes

Whether objectives are determined logically according to analyses of learning processes and content (taxonomies), the structure of disciplines (Bruner), or behavior (Mager) or whether they are based on the theoretical approaches of Guilford (intellect) or Gagné (learning), they must be defined clearly. Vopni (*5*) uses a series of question to test for clarity:

> Who is to exhibit the behavior? What action is the learner expected to perform? What is the situation that stimulates the learner's performance? What object is being acted upon or

interacted with? What constitutes the set of acceptable responses? What special constraints or restrictions or limitations, such as time or materials available, are imposed?

In the past, a typical statement of an objective might have read: To develop an understanding of principles of structural analysis. Since this objective neither defines nor limits behavior, *any* of the following could show that the learner understands the process:

1. The learner says, "Man, that's beautiful. Look at those splits!"
2. The learner successfully completes his syllabication practice sheet well in advance of his peers.
3. The learner nods in agreement as another student performs syllabication correctly at the chalkboard.

Today, the reader of curricular objectives expects to find certain information within the statement that will tell him what the learner is doing when he is "understanding" principles of structural analysis. The statement might read: The learner must be able to divide into syllables a list of unfamiliar words with at least 90 percent accuracy.

Ideally, then, the clearly stated goal is one that contains the instructional intent of the person writing or selecting the objectives. In addition, it will define the *criterion* of acceptable performance.

Many schools are presently reformulating the objectives of their reading programs. As a first step, educators often divide the reading curriculum into basic components which can be placed on a continuum of difficulty from the easiest skills to the most complex. As a result, more teachers than formerly are beginning to know the structure of the entire reading program from its introduction to its conclusion. When reading skills are translated into specifically stated objectives, teachers are able, as perhaps never before, to place children at levels where they can operate effectively. Teachers also are more knowledgeable about the skills which their pupils can or cannot apply, particularly if the teachers use informal test items based upon specific reading objectives from the continuum. In other words, a *criterion reference* is helpful—can the child do the task? Thus, the child's actual performance provides evaluative information which can lead to appropriate prescriptions.

Obtaining Evidence and Making Interpretations

Two major acts are involved in accomplishing the evaluation process: *description* and *judgment*. While many evaluators of the past chose not to judge, it is more than likely that judgments will be found in greater quantity in evaluation reports of the future.

Whenever certain data are gathered from a number of sources and by a variety of means, three bodies of information should be distinguishable. Stake (*3*) identifies these as *antecedent conditions, transaction,* and *outcome data.* An *antecedent condition* exists prior to instruction, whereas *transactions* are the interactions of pupils with relevant and multimedia during the educational experience. *Outcomes,* traditionally receivers of the lion's share of attention in formal appraisals, consist of the abilities, attitudes, and aspirations of students subsequent to the learning period. The format of Stake's 12-celled matrix (*4*) may be helpful in organizing the collection of descriptive (intents and observations) and judgmental data (general standards of quality or judgments specific to the program). In preparing a record of this nature, the evaluator notes what educators intend to accomplish, what observers actually see taking place in the classroom, what school and community expectations of the program are, and what the evaluator judges the program to be.

The following illustration from an upper elementary reading class demonstrates how data can be entered, beginning with *antecedents* or entry behaviors and progressing down each column: Knowing that 1) on Monday a class of students will be assigned materials which employ a variety of techniques to sway readers' opinions and that the instructor intends 2) to lead a discussion on the topic on Tuesday, the instructor indicates 3) what students should be able to do on Wednesday, partly by locating additional samples from magazine advertisements and by written summaries of several types of propaganda techniques. He notes that 4) some pupils are absent on Tuesday, that 5) the class period concluded before all techniques had been discussed, and that 6) only about one-half of the class had located appropriate magazine ads on Wednesday and that about one-half of the group had prepared accurate descriptions of these techniques. In general, the instructor anticipates 7) some absences but

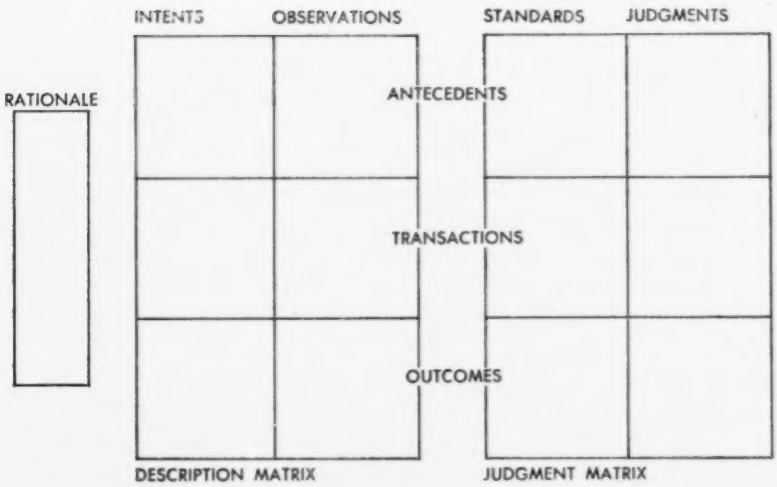

FIGURE 1. A layout of statements and data to be collected by the evaluator of an educational program (4).

that the work will be completed by the following class period; he expects 8) his discussion to be clear enough for perhaps 80 percent of the class to grasp the concepts presented; and he knows that 9) his team teachers expect about two pupils in ten to fail to understand concepts of similar difficulty. By his own judgment 10) the assigned materials lacked sufficient clarity; the students showed interest in the discussion 11) through their questions and comments; and his teacher aide who read the summaries and accompanying illustrative student-collected materials said that 12) a large number failed to distinguish between three types of propaganda devices—testimonials, glittering generalities, and card-stacking.

Naturally, it is not necessary to record data in such detail, but a frequent "thinking through" of these steps should aid in evaluating how well certain goals are being achieved. In addition, such an analysis could help in discovering relationships which in turn could enable educators to improve future class sessions and programs.

A second example of the evaluative process will illustrate how a group of primary teachers is engaging in an early assessment program

(*2*). The following description shows how the McKinley Complex schools are implementing several levels of continuous evaluation.

The first level includes all children at all times. From "*on-the-spot*" diagnosis, teachers learn whether children succeed as they perform specific reading tasks. If pupils are still not successful after teachers have adjusted instruction, an analysis of teaching procedures should take place. And if new strategies prove ineffective, teachers can move to the level of *formal diagnosis* during which standardized diagnostic reading tests and informal inventories may be administered. Looking at the information they have gained from these instruments, teachers may be able to formulate hypotheses based upon pupil patterns of reading behavior. If further study of an individual's difficulties is needed, teachers can proceed to *detail diagnosis* which may include the testing of specific skills in greater depth and, if required, referral to agencies for clinical services.

On-the-spot diagnosis involves continuous checking of pupil's daily reading activities. From these activities important information can be gained about pupil abilities, perhaps by a child's explanations in "show and tell," his use of language in oral conversations, his written responses to visual and auditory stimuli, his eye-hand coordination skills, and his performance during oral and silent reading. Through such observation, the teacher can understand many of the correlates of reading achievement and thereby be able to develop tentative hypotheses regarding the needs of each pupil. If, for example, visual discrimination appears to be a problem, the use of a checklist of symptomatic visual difficulties may be helpful in determining whether the child sees clearly enough for accurate discriminations among letters and/or words. Checklists and rating scales facilitate precise observations by allowing the teacher to focus on specific behaviors. On the other hand, if a child's show-and-tell experience indicates his difficulty in relating a sequence of events, the teacher can note the problem and plan practice opportunities to aid the pupil in expressing ideas in logical order. Obviously, no one evaluative source, such as teacher observations, should be relied upon to the exclusion of others which could be more productive in supplying relevant data.

Formal diagnosis is initiated when observational procedures fail

to bring to light evidence enough for individualizing instruction. At that time, the teacher can check on specific skills, such as sight recognition vocabulary, knowledge and application of various phonics techniques, and performance on oral and silent reading inventories. Accurate interpretations at each step of the process are essential. In discovering Ed's poor comprehension at the beginning of the year, his teacher could have assumed that he had poor recall of details. Instead, when she administered an informal inventory she found excellent recall of story details in company with accurate word analysis techniques. Even at the easiest levels, however, Ed lacked ability to state main ideas and to make inferences. His teacher suspected that his previous experience in comprehension practice might have overemphasized details at the expense of main ideas, since others in her class appeared to have similar problems. She then checked the skills continuum in this area and decided to test his ability to locate main ideas when several responses were given. Ed could not select the best statement from among three others; he appeared to be so concerned with facts that he could not see "the forest because of the trees." When the teacher took him back to a lower level on the continuum, gaining main ideas from pictures, Ed experienced success. The following schema illustrates her procedures.

	"On-the-Spot"
Observed Behavior	Poor comprehension (inability to recall details)
	Formal Diagnosis
Limited Testing (Informal reading inventory)	Excellent recall of details; accurate word analysis skills; failure to state main ideas in his own words; failure to make inferences
Criterion reference (skills continuum)	Inability to locate main ideas when several were suggested Success in stating main ideas from pictures (Gates-Peardon Practice Exercises in Reading, Type A, and paragraphs from story and content subject material)
	Formulation of Hypothesis
Individualization of Instruction	Use of pictures to find central theme; use of story material for practice in selecting the best statement of main ideas from four—followed by discussion of

why the others were not appropriate; use of content material to identify sentences which contained key words and ideas; stating main ideas and comparison with teacher or workbook model; supplying endings for stories; predicting what would happen next in fictional content and science materials; determining mood, setting, weather, season, in story material when these items were not stated directly.

Had the teacher provided practice in noting important details according to her initial limited knowledge of this boy's problems, Ed's comprehension difficulties could have multiplied. Individualization of instruction requires evaluation of pupil weaknesses, followed by accurate interpretations of findings and observations.

Although further diagnosis was not indicated in Ed's case, teachers occasionally need to go the next level of formal diagnosis. *Detailed* diagnosis explores the ways in which a child learns—the ways in which his psychosensory system works best. Does he perceive, remember, and interpret what he hears, sees, or feels? To answer these questions, investigations may be initiated to assess three types of learning: 1) *intrasensory*—learning requiring only one system such as vision or hearing; 2) *intersensory*—learning requiring two or more but not all systems; and 3) *integrative*—learning requiring all systems functioning as a unit. The study of each learning style includes a variety of measures of perception, memory, symbolization, and others. By this process it is possible to determine by which modality a child should be taught. Knowing this fact, the teacher can select methods which are appropriate for the child's learning pattern.

Conclusion

Today, within the profession, there are those individuals who are devoted to analyzing the tasks involved in the successful completion of a designated behavior. Educators can no longer avoid this new area of specialization and technology; school people must think in more analytical terms as they develop and use curriculm study guides.

It follows that evaluation must be realistic and pragmatic, as

well as somewhat idealistic. The total waste of resources that accompanies some elaborate schemes which are excessively time-consuming and virtually impossible to perform should be corrected. Similarly, those projects which have such limited evaluation plans that little of value is forthcoming for the planning of individualized instruction should be improved qualitatively and quantitatively. Hopefully, some of the strategies suggested in this paper can be utilized.

REFERENCES

1. Dale, Edgar. "Vocabulary Measurement: Techniques and Major Findings," *Elementary English,* 42 (December 1965), 895-901, 948.
2. Hawaii State Department of Education. *Individualizing Reading Instruction through a Diagnostic Curriculum.* Honolulu: Office of Instructional Services, 1969.
3. Stake, Robert E. "The Countenance of Educational Evaluation," *Teachers College Record,* 68 (April 1967), 523-540.
4. Stake, Robert E. "The Countenance of Education Evaluation," *Teachers College Record,* 68 (April 1967), 529.
5. Vopni, Sylvia. "What Is Evaluation?" *Educational Horizons,* 47 (Winter 1968-1969), 75-81.

The Diagnostic Teaching of Reading

WANDA GALE BREEDLOVE
Educational Service Center
Lexington, South Carolina

"HOW WILL we know where we're going if we don't know where we are?" The census slogan well-known to an estimated two hundred million Americans suggests the importance of evaluation *before* planning. *Evaluation* and *planning* are words common to education but often divorced from each other. Planning is requested at the beginning of instruction and all too commonly evaluation is required only at the bottom of a lesson plan, at the conclusion of a project, or at the end of the term. In contrast, diagnostic teaching is the interrelation of evaluation and planning contiguous with instruction.

Textbooks on the teaching of reading, in particular, emphasize planning and evaluation prior to, during, and after instruction to determine the reading status of pupils, to diagnose specific strengths and weaknesses, and to assess pupil progress (*12*). A most useful strategy for teaching diagnostically often cited and discussed in reading circles is a test-teach-test-teach cycle that is open-end and continuous (*5*).

Test-Teach Cycle

This brand of diagnostic teaching refers to some form of teacher observation and testing that involves an initial "look" at the child on a specific learning task to determine what, if any, of the required skills he does not now know.

Evaluation initiated before teaching gives the teacher direction. It not only assesses where the students now are but also indicates where the students are not—the *are now* being those skills and abilities that the student exhibits under the testing conditions and the *are not now* those skills and abilities that he is unable to perform (*3*).

The teacher can then plan a program intended to assist the child

in learning or applying these skills and abilities. A check-up or post-test on the same task assures the child and the teacher of mastery or the necessity to develop an alternative teaching strategy. Mastery necessitates setting up a pretest for the next skill or ability. Lack of mastery presents a problem-solving task to the teacher and the learner. The cycle of test-teach-test-teach- is self-perpetuating—continuous. It leads from one test to the next and one plan to an alternative on an individual basis through all skills, methods, and materials. To elaborate on the test-teach pattern, consider the role of the learner and the role of the teacher in diagnostic teaching.

Reading: An Empirical Experience

Reading is an empirical experience. The learner is continually testing his perception of the printed page with his stored past experiences and making predictions. Each response is an experiment—a test run to determine whether what he reads reflects the print on the page. This experimentation is based on the learner's experiences and will be interpreted in light of his peculiar view of himself and the reading act.

Reading, whether interpreted as an act of decoding, encoding, or both follows a pattern of formulating tests and evaluating results, as do other dimensions of learning—as does teaching! Further, neurophysiological and linguistic theory would appear to reinforce the essential character of testing and evaluation for the diagnostic teaching of reading if, as Miller, Gallanter, and Pribram (*8*) submit, testing and evaluation are integral functions of the brain in processing information and dictating action in the learning process. Testing every operation he performs is the only way a learner can confirm that he is reading c-a-r as *car* or indicate to himself that he needs to modify his information, experience and/or environment, or the way he manipulates the information so that he will be reading c-a-t as *cat* and not *car.*

Role of the Learner

Testing is vital to the learner. The results of the test, correct or incorrect, are important both to the learner and to the teacher. A

significant role of the teacher is to supplement a student's inadequacy in testing his own performance. The teacher can temporarily assume a testing facility for the child. Snygg (*10*) considers the implications of the value of error recognition to the student in his cognitive field theory of learning when he suggests, ". . . learning does not take place unless the learner finds that he has made a mistake. . . ." There is no reason for the learner to respond differently unless a response does not work. If a response does not work, then the learner must try to combine his information in other ways, reexamine the environment for more clues, and sort the experiences he has had for similar situations; in short, he must change his way of thinking about the particular problem. An error alerts the student that there is an element in the situation that is different from what he has supposed. Predictions, operations, and evaluation follow.

The reader who corrects his oral reading on the basis of sentence sense is running one such trial when he reads, "Sam is a cool car—I mean, cat." This reader has set an objective—meaning. He proceeds through the material and might well read, "The villagers journeyed west" as "The villagers traveled west" with no oral correction or recognition of miscalling a word. The sentence passed a meaning objective; none for word recognition was in effect. Learners determine their own objectives and run tests accordingly. Sometimes learners agree with those of the teacher; often they do not. It is obvious that objectives based on individual perceptions differ for individuals. Certainly, the more closely the teacher can set objectives with students or take existing student objectives to achieve her own criteria of student performance the more likely specified learning is to occur, especially if learning is the individual *becoming* through his own processes of experiment.

A focus on learning implies change—change on an individual basis and change that develops after some difficulty in meeting a situation. The optimum level of difficulty proposed by emphasis on error recognition would be "one which allows the student to win success after difficulty"(*10*). The ensuing change can and often has been measured in terms of specific behaviors. But in addition there is now measurable evidence that the process is more important than the product. Learning expresses itself experimentally as evidenced by the increased weight of the brain of laboratory animals (*6*). Learn-

ng as change is structural and chemical change and is prior to behavioral change. It is change within as well as without.

As the learner formulates and modifies predictions, he learns. Imagine a young child who predicts that *ed* forms the past tense of verbs. He has heard others use *ed* in statements, like "Mommy walk*ed* to the neighbors" and "Brother stay*ed* at Grandma's." The child generates, "I went*ed* with Daddy" by forming a word he has never spoken before. This act cannot be imitation. It is a more important process. He has predicted that if he did something in the past, *ed* will express what he wants to say about it in the present. But this prediction, no matter how dependable, will not always work. Other ways of sorting information and predicting must be found so that the learner can exercise control over his environment. Mother responds, "You *went* with Daddy yesterday." Mother provides a model for what the child wants to be. She "teaches" by example and provides constant feedback for the child, by pointing out incongruities that the child has not had the experience to recognize himself and by expanding his restricted information.

Role of the Teacher

Now to refine the *role* of the teacher in the diagnostic teaching process, the teacher can serve much the same role in learning as a parent does in language acquisition. The teacher, too, can serve as model and reactor, but there are unique ways in which she can facilitate learning. A brief review of the learner and the learning process illuminates the diagnostic teacher's role. Learning is becoming. Applied to the teacher, this statement means that a teacher need not have all the answers. She, too, is a learner. Another important consideration is the value of error recognition. The teacher may learn most herself when her usual answers to student learning difficulties are least effective. It is then that she searches for new ideas and alternate ways of presenting information. It gives the teacher an opportunity to increase the weight of her own brain, to work through strategies, to exercise her mental prowess. Teacher activity of this kind is problem solving. Personally and professionally, the teacher in such a scheme is always becoming her best self.

The teacher as learner values student errors as well as her own. Rosenburg makes student errors the primary source of information for adjusting teaching methods (*9*). If the teacher analyzes the results of tests, she can develop a plan of processes, procedures, and materials to "try out" with each student based on his specific errors or patterns of errors. The teacher's plan might best be thought of as "testing" hypotheses based on a comparison of information collected from student responses during paper and pencil testing and informal observation with how she presented materials and information to other students—a process of making the "best guesses" she can to help her solve the teaching problem of a particular student on a specific skill under specified conditions, unique to any teaching problem she has ever encountered before.

Suggestions for the Classroom

The teacher may well puzzle, "Diagnostic teaching sounds good, but does it have any practical application for my classroom of thirty students? Is there any way to work it in with basal lessons? Isn't diagnostic teaching something like individualized instruction? What materials can I use? How can I find time? Will it really make any difference in the reading level of my students?"

Frankly and honestly, there are no conclusive answers to these queries. The individual teacher-learner is probably the only one who can attempt an answer. Even then, the answer will only be a tentative one—a working hypothesis until tried with learners, modified, and evaluated by teacher and student.

A response to these questions is, however, possible in terms of the search that teachers have made in their classrooms and research that has been conducted in other learning laboratories. I would like to submit suggestions for the classroom in three areas introduced earlier in this discussion: a student's current reading status, specific strengths and weaknesses, and progress.

Initial testing and evaluation screen general student performance and direct additional testing. In a seventh grade spelling class the teacher assigns Unit 32, "The Consonant Sound *N*." Students work on assignments in the unit during the week, and on Friday

the teacher dictates words from the unit. The expedient of giving the same unit test on Monday can eliminate those students who already spell the words in the unit and thereby free them to work on spelling words that they cannot now spell. Those who cannot spell the words on Monday, but by studying in the book can do so by Friday, continue the routine and demonstrate learning-gain success after difficulty. Those students that cannot spell the words on Monday and cannot spell them on Friday, try studying words devised from less difficult lists or from a different way of working with the same words in terms of length of time and alternative exercises. Granted the teacher has not yet located specific ways of working with each student on specific skills; she has begun to place students on instructional levels and through additional evaluation during the course of classroom assignments will locate patterns of errors. This first rough testing in the subject helps the teacher to determine instructional level, much as reading out of a graded basal reader or subject matter textbook in a group informal reading inventory alerts the teacher to groups of children that can proceed in the materials at their grade level, as well as to groups of children that require testing in graded books at lower levels of difficulty, and to others that can perform reading tasks far above grade level.

Materials, Procedures, Processes

As the teacher tests students in multigraded texts, she discovers one student has no way of attacking words, another can sound out letters in words but is unable to blend them, a number of students can answer comprehension questions when the teacher reads a selection to them but cannot answer questions from their own reading, or that many of her students cannot answer questions requiring reorganization of ideas. Evaluation of this kind alerts the teacher to areas of instruction appropriate for individuals, small groups, and the whole class.

The search begins for materials on the reading levels of some pupils, word analysis work sheets, tapes of phonic blending exercises, and reading paragraphs that ask for reordering sequence of events. Some of these materials will serve as further tests on particular stu-

dent problems, and the teacher will predict other ways of working from these additional pupil responses. A student may not respond to a piece of material at all. This lack of response may lead the teacher to choose other reading selections or to assist the pupil in developing his own materials.

As the teacher views pupils more closely for specific strengths and weaknesses, she is looking for an every-pupil-response. Essentially the teacher directs each individual to materials and with methods that result from predictions made from individual responses. If in a classroom of thirty pupils students are to respond to different items, the teacher must incorporate into the program some form of self-checking material. Items on teacher-developed reading kits require individual pupil responses and self-checking. Teachers organize workbook exercises, stories, and pictures into kits corresponding to the order of skills presenting in basal materials or on scope and sequence charts coded at a number of difficulty levels. In his book, *Materials for Remedial Reading*, Gilliland (*4*) offers suggestions for arranging these materials to include varied presentations of skills from many different texts. The workbook sheets serve as initial tests of skills. Students proceed to easier or more difficult exercises or to another skill altogether on the basis of their responses. Answer keys mounted to the back of these materials encourage self-correction. Plastic film placed over the pages and marked with a grease pencil or crayon that can be wiped clean render the materials reusable. The teacher can collect and with the assistance of her pupils construct extensive files of readily available exercises on every conceivable skill at all levels of difficulty.

In addition to the representative materials cited certain teaching procedures expand the usefulness of testing-teaching patterns in the classroom. One of these techniques, the construction of instructional objectives, assists the teacher in formulating teaching goals with built-in tests and established levels of mastery. To clarify this relationship, examine the following objective stated in behavioral terms: "The learner will be able to name the letters of the alphabet when presented in random order with 95% accuracy." Teachers of beginning reading might well want children to recognize the letters of the alphabet. The behavioral objective emphasizes the test for

learning, observation of the learner, and evaluation of the product. Evaluation is built into teaching in the planning stage of instruction (7).

Numbers of objectives are operating in the classroom for groups and individuals simultaneously. Checklists are an effective way of cataloging objectives for easy access. Structural analysis skills, visual symptoms, reading interests, and language patterns are a few of an infinite variety of observations that are possible. During instruction constant checks can be conducted for each pupil on a variety of skills and abilities. A favorite format is one that includes the names of all pupils on the same sheet to aid in the comparison and contrast of pupils that exhibit similar patterns for grouping purposes and team learning or highly individual responses for special attention.

As evaluation continues in the teaching strategy, more and more instruction will be directed for individual pupils because unique characteristics of pupils once recorded will dictate the formulation of widely differing hypotheses about the approaches and rates and materials in each experimental process. An individual conference provides an appropriate setting for guiding individual study. A brief discussion with each student to develop learning prescriptions is central to testing student and teacher predictions for their mutual evaluation of the learning process. Process, after all, is the chief concern—a problem-solving attitude developed by solving problems.

A problem to solve is one that the learner recognizes. He compares *what is* with his conception of *what ought* to be and solves the problem by "filling in the gap." Filling in the gap may be accomplished by one of a number of heuristic devices, such as relating the unknown to the known, restating the problem in other words, solving a related problem, decomposing the problem into simpler problems, working backward, or omitting certain details.

The teacher translates her objectives into student tasks. She offers feedback that confirms a match or mismatch. Based on a comparison of his response with what is expected, she tests heuristic prescriptions in the form of oral discussion questions or written exercises to lead the student to a learning experiment. Teacher preinstruction for the experiment consists of organizing items and stimuli in order to direct student "tests."

Given the previously stated objective of naming the letters of the alphabet, a learning experiment develops. The learner predicts that the symbol *j* on the first flashcard presented represents the oral "jay." As he continues naming the letters and confirming his "educated guesses," the teacher notes a pattern of misnamed letters *b*, *p*, and *d* on a checklist. She hypothesizes that an important element in this pattern of errors may be an inability to distinguish these symbols visually. The teacher decomposes the original problem into simpler problems. She selects a paragraph and asks the pupil, along with another who is misnaming *q* and *p*, to circle *p* with a red crayon each time it occurs in a paragraph. She gives each pupil a sample of the letter. She then asks both pupils to complete the same activity for another troublesome letter with a blue crayon. When completed, each student tests his associations with those of the other. One of the students is unable to distinguish between the various forms of the letters *q* and *p* visually. The teacher reevaluates and directs the student to cut out *p*'s from a magazine. The student that completed the visual task requires a different prediction. "Say the name of the letter I say after me," requests the teacher. She flashes the symbol *b* and says "bee." The student responds, "bee." He predicts "bee" when *p* is shown as well. Predictions and tests continue using various auditory materials and questions until the student organizes the auditory information in these related problems into a scheme for naming the letters in random order.

Student hypothesis testing and mutual evaluation in small groups is equally practicable for developing comprehension skills. In a most intriguing discussion of *Teaching Critical Reading at the Primary Level,* Stauffer (*11*) encourages students to predict story endings, test their predictions one with another, and to change predictions as the story unfolds. It is this willingness to make "best guesses" and to continually evaluate ideals and attempt other responses in the light of a problem that best characterizes the student under the influence of the diagnostic teacher. Listening to students discuss their own problems and how they attempt to solve them gives clues about the individual thinking-testing process and affords the teacher unusual opportunities for observing current mental operations and suggesting others (*8, 12*).

For the teacher interested in developing diagnostic teaching abilities, teacher-training materials are available. One, a test of problem-solving facility, gives practice in evaluating pupils and predicting instructional tactics (*1*). Simulation exercises produced by Della-Piana and associates (*2*) involve comparing student reading responses with possible teaching strategies to formulate instructional hypotheses.

The foregoing discussion calls for a diagnostic search of the total learning field for materials, techniques, and processes. It asserts the need for an experimental attitude in teacher and learner to constantly evaluate change and the tentative answers that change brings. From the variations of the print on the page to the firing of electrical impulses in the brain, the learning payoff is the potential for the learner and the teacher-learner to be engaged in the process of evaluating, predicting, acting, and becoming more able to learn in the process.

REFERENCES

1. Burnett, R. W. "The Diagnostic Proficiency of Teachers of Reading," *Reading Teacher*, 16 (January 1963), 229-239.
2. Della-Piana, Gabriel, Betty Jo Jensen, and Everett Murdock. "New Directions for Informal Reading Assessment," in William K. Durr (Ed.), *Reading Difficulties: Diagnosis, Correction, and Remediation*. Newark, Delaware: International Reading Association, 1970, 127-132.
3. Freidman, Miles, et al. "Readiness and Instruction: Individual Diagnosis and Treatment," in Edith Grothberg (Ed.), *Critical Issues in Research Related to Disadvantaged Children*. New Jersey: Educational Testing Service, 1969.
4. Gilliland, Hap. *Materials for Remedial Reading and Their Use*. Billings: Montana Reading Clinic, 1968, 97-104.
5. Harris, Albert J. *How to Increase Reading Ability*. New York: David McKay, 1968.
6. Hilgard, Ernest R., and Gordon H. Bower. *Theories of Learning*. New York: Appleton-Century-Crofts, 1966.
7. Mager, Robert F. *Preparing Instructional Objectives*. Palo Alto, California: Fearon Publishers, 1962.
8. Miller, George A., Eugene Galanter, and Karl H. Pribram. *Plans and the Structure of Behavior*. New York: Holt, Rinehart and Winston, 1960.
9. Rosenburg, Marshall B. *Diagnostic Teaching*. Seattle: Special Child Publications, 1968.

10. Snygg, Donald. "A Cognitive Field Theory of Learning," *Association of Supervision and Curriculum Development Yearbook*, 1966.
11. Stauffer, Russell G., and Ronald Cramer. *Teaching Critical Reading at the Primary Level*, Reading Aids Series. Newark, Delaware: International Reading Association, 1968.
12. Strang, Ruth. *Reading Diagnosis and Remediation*, ERIC/CRIER/IRA Reading Review Series. Newark, Delaware: International Reading Association, 1968.

How the Classroom Teacher Can Use A Knowledge of Tests and Measurements

MARVIN GLOCK
Cornell University

THERE ARE many facets of measurement that a teacher must understand to use tests effectively. To discuss this topic thoroughly requires much more space than limitations here allow, so we must choose from the alternatives open. Two of these are quite different: one, to list a number of principles; the other, to select two or three basic concerns in measurement and illustrate their importance for classroom teachers. I have chosen the latter alternative in the belief that it is more likely to be of greater practical value. Discussion, therefore, will be limited to test validity, reliability, and the problems in measuring gains in achievement.

Validity

Does the test measure what it is intended to measure? One would be unwise to assume that the name or title of a test tells what it measures. If you examine various tests of "reading comprehension" you will find that certain tests require the pupil to determine the main idea of a paragraph; others demand only the retrieval of literal meaning; and a few ask the reader to discern the intent and mood of the author. Tests of reading rate vary in time from 60 seconds to much longer periods of reading. Some give credit for speed, even if many questions are not answered correctly; others give no credit for rate unless comprehension is assured. Some vocabulary tests are constructed of items listing a word with several possible synonyms from which to select the correct answer. In others, the examiner reads a sentence and the pupil is required to select one of several words to complete the sentence. In still a third type, the pupil is presented with a sentence and an underlined word. He responds by marking

one of several possible synonyms. There are also vocabulary tests requiring a pupil to define a word and to use it in a sentence. It is very possible that some individuals will perform better on one type of test for comprehension, rate, or vocabulary than on others. Yet, all may have identical titles. Which tests are valid? Which tests are measuring what you as a teacher want them to measure?

In comprehension tests we find that some instruments are very limited in what they measure. Davis (*1*) lists eight skills that determine good reading comprehension: 1) recalling word meanings; 2) drawing inferences about the meaning of a word from context; 3) finding answers to questions answered explicitly or in paraphrase. 4) weaving together the ideas in the content; 5) recognizing a writer's purpose, attitude, tone, and mood; 6) identifying a writer's techniques; 7) following the structure of a passage; and 8) drawing inferences from content. Naturally, as Davis suggests, some of these skills are more important than others. After careful study and analysis of a particular test, however, we found questions on only two skills: recalling word meanings and finding answers to questions answered explicitly in the passage. We would rightly conclude that this test had low content validity. It would not be a valid test for the purpose of measuring total comprehension.

A number of publishers provide an analysis chart for their tests that indicates what is believed to be the content, type of material, or skill being tested by each item. This information is helpful. The procedure cannot, however, replace the need for the teacher to take the test himself—to expose himself to the tasks presented. He can then check the chart against his own judgment. It is obvious that different types of questions require pupils to respond in different ways and this, along with content, determines what the test is measuring.

There is another factor that determines what the test is measuring: the care with which the questions have been constructed and tested. A poor question may enable a pupil to select the correct answer: for example, not by determining the main idea of a paragraph but by matching a word in a question choice with the identical word in the passage.

Going to the park one day on his way to school, Bill stopped and watched them paint the new pavillion. The bright yellow color sparkled in the sunlight.

What color was the pavillion painted?

1. Red
2. Yellow
3. White
4. Green

Poor questions also allow the pupil to eliminate implausible answers and select the correct one without the comprehension intended by the author. For example, one test includes the following item to be answered after reading a selection.

The chief factor limiting the amount of land for cultivation is:

1. rugged peaks
2. climate
3. irregular coast line
4. poor farming methods

The pupil does not have to understand the passage to answer the question correctly; he could even choose the correct answer without reading the passage. Common sense dictates that none of the last three choices limits the amount of land for tillage; only the first choice could possibly be correct. Poorly constructed tests allowing test-wise pupils to respond correctly by means of irrelevant cues are not measuring what was intended.

Another factor bearing on test validity is the adequacy with which it samples reading skills or knowledge in vocabulary tests. Test designers are limited in the number of responses they can ask a pupil to make. For example, in a vocabulary test only a few of the words that a pupil might be expected to know are included. Another vocabulary test might present an entirely different list of words. The manner in which these lists are selected will determine how well the test depicts vocabulary development. Some lists include general vocabulary; others may be loaded with scientific terminology. Various kinds of bias can exist.

The care with which the test is administered can influence validity. Scores may be consistently too high or too low because of administrative procedures. In standardized tests instructions must

be carefully followed giving no more and no less help than is specified. Time limits must be adhered to. Room conditions and seating arrangements should provide for optimal performance. Interruptions and other distractions must be eliminated. No teacher should administer any standardized test without first carefully reading through all instructions and underlining the time limits, in color preferably. It is assumed, of course, that all pupils will answer the same questions.

Reliability

Another important quality of a good test is adequate reliability. No psychological test can measure as precisely as a foot rule or even a household scale. On the other hand, good tests do reflect the quality of *reliability*. When we speak of reliability in a person, we imply veracity and complete dependability on what the person does and says. Reliability in a test implies that it is consistent in what it measures. A very reliable test may not be "telling the truth," but it continues to report the same falsehood quite accurately. For example, if we administer a paper and pencil verbal intelligence test to a child who cannot read, the child would invariably make a low score, an indication of low intelligence. The test might have high reliability, but it would not be measuring intelligence. Rather, it would be revealing a child's inability to read. It would not be telling us the truth about his intelligence.

Consider another illustration. Suppose we measure height with a yardstick that is marked off only in feet. There are only two marks on the stick. It is obvious that the measurement of height would probably be less accurate than if the stick were calibrated to 1/16 of an inch. Getting two measurements alike with the rough markings is most unlikely. Such a poor instrument would give inconsistent measurements; it is unreliable. Likewise, if we are to have confidence in a test, the score must be attained by careful measurement.

There are certain factors that determine test reliability. First, we must have a number of samples of a pupil's performance on a task. We certainly do not judge a batter's skill by one time at bat. His batting average is determined over a series of performances at the plate. Neither do we judge a pupil's vocabulary effectively by

giving him a word or two to define, nor even ten or twenty words. A pupil must give the meaning of a great many words before we are able to get a precise measure of his ability. By the same token, we will be unable to have confidence in a reading comprehension test score unless the child answers a considerable number of comprehension questions. The more questions—that is, the more samples—the more likely the test is to be reliable, other things being equal.

That's the reason part scores on a test are so often suspect. Each part is a small test; only a few questions yield a part score. Sampling with such a limited number of tasks is inadequate to give a reliable score.

A test's reliability also depends upon accurate scoring. Scoring is not difficult with objective type tests; there is an obvious right or wrong answer. Scoring becomes more of a problem when pupils write out their answers to a question. There are specific procedures to improve scoring reliability in these instances.

Of course, there are available data provided with all good tests for the user to determine the adequacy of the test's reliability. In general, for an individual pupil's test score to be reliable enough for proper interpretation, there should be at least one half hour of testing time with a minimum of forty to fifty questions. A reading test with items demanding complex and critical thinking will need more questions for optimum reliability than does a test requiring mastery of literal meaning and factual information. Insufficient length prevents subtests of ten to fifteen questions from attaining adequate reliability. Also, a test that is designed for testing pupils in a wide range of grades, e.g. grade three through grade twelve, may have only a few questions that are suitable—not too easy nor too difficult—for the children in any one grade, thus resulting in a very short test for each pupil.

We have been discussing reliability chiefly in terms of standardized tests. How is this information related to the short, teacher-made, classroom tests? Well, certainly the teacher doesn't want to make important decisions on the basis of test performance involving only four or five questions. Over a period of time, however, if a teacher is consistent about administering these short tests, he will build up a considerable number of questions—in effect a long test—whose

reliability will most likely be adequate to aid with the help of other information in making valid judgments about each pupil.

But no test is perfectly reliable. For the practical situation there is always an error of measurement in a test score. When a pupil earns a score on a test we never know, therefore, whether it is higher or lower than deserved. Many of the better tests use a special system to help the teacher interpret test scores. Raw scores are changed to percentiles, and a score is reported as falling within a percentile band. For example, in Figure 1 if Mary's converted score in reading was 300, the band between the 60th to the 82nd percentiles would be an indication of the possible error. However, we could say with reasonable certainty that her score was better than 60 to 82 percent

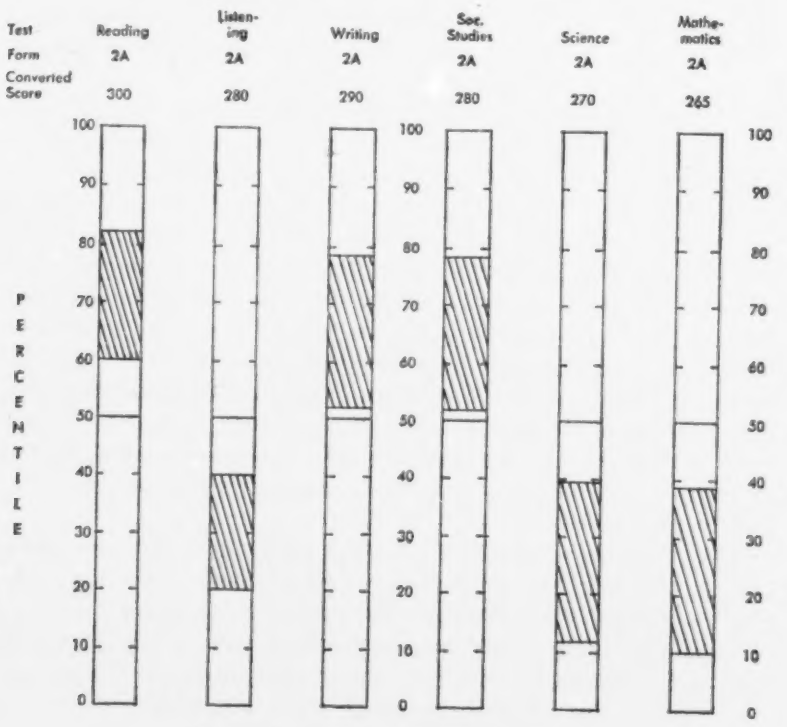

FIGURE 1. Pupil Profile Chart—STEP

of the standardization group. Also, since this is a battery of tests, we could draw some conclusions about Mary's comparative performance in several subject areas. The percentile bands of reading and writing overlap. We could not, therefore, conclude with assurance that her reading ability was greater than the score in writing because of the measurement error in both scores. However, it would be reasonable to conclude that her reading score does represent superiority when compared with listening, science, and mathematics because there is no overlapping. The teacher who realizes that he must interpret test scores with great care because none is free from error, will also muster as much additional information as possible before making important judgments and decisions about children.

Measuring Gains in Achievement

One purpose of tests is to determine the progress of pupils. Initial and final scores on standardized achievement tests are often ascertained. But when a test measuring skills such as reading, arithmetic, and writing—skills which develop more or less continuously —is administered at the beginning and end of the school year, the average score is almost certain to rise. However, if we look at the following scores in Figure 2, we note a phenomenon that could be embarrassing. Note the difference in gains among the various groups. We find that the lowest group has gained the most; the highest group has gained the least; and the middle groups have gains in between. Some of the high group would appear to have lost achievement in other subject areas. How can we account for this state of affairs? Have the lowest students actually learned more because instruction was pitched at their level while, in the meantime, the high achievers just marked time?

This assumption might seem reasonable if in this well-known study the same phenomenon had not occurred with interest inventories, inventories of beliefs, and problems in human relations. Even in the affective realm of attitudes, values, and personal-social adjustment those who made the lowest scores gained the most on a second administration while those making the highest score gained the least or made a lower score.

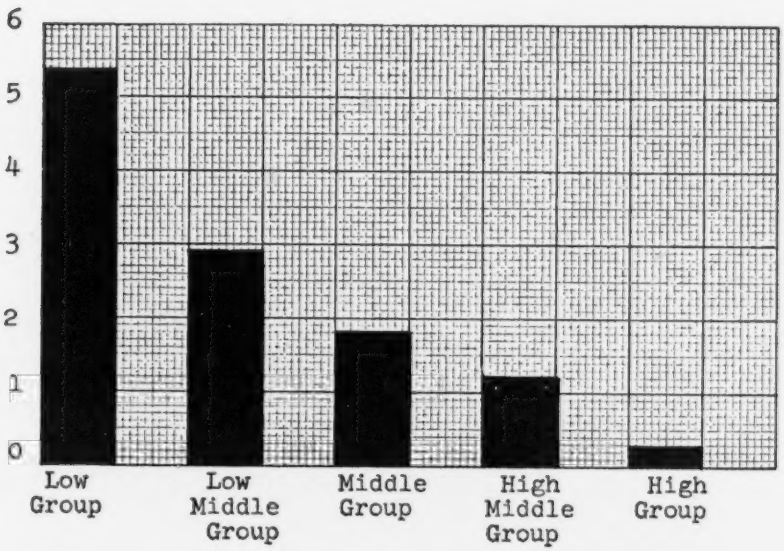

FIGURE 2. Average gains of students on post-tests, classified according to pretest standing (*2*).

If we are not to conclude that students with initial high scores change their behavior very little, if at all, while those with low scores learn a great deal, then we must look to the construction of tests or to the testing repetition for an answer.

One answer to the problem is that tests may have been too easy for the better pupils so that they very nearly answered all of the questions on the first testing. There would be little if any opportunity for improvement because another form of the test would be of the same level of difficulty.

Another explanation of why lower achievers make greater test score gains than higher achievers is the phenomenon of "regression." In 1889, Galton (*3*), an Englishman, reported that short parents tended to have children who were taller than they and tall parents tended to have offspring who were shorter than they. He stated:

> However paradoxical it may appear at first sight, it is theoretically a necessary fact, and one that is clearly confirmed by

> observation, that the stature of the adult offspring must, on the whole, be more mediocre than the stature of their parents (*4*).

Galton called this tendency the law of regression and related it to various hereditary traits. It can be easily explained. There is not a perfect correlation between parents' and children's heights. Therefore, the children of tall parents will be shorter and closer to the mean. If there was a perfect correlation, children would be as tall as their parents. They cannot be taller since their parents are already at the extreme end of the distribution. The same kind of reasoning holds for the relationship between short parents and their children.

Regression is observable whenever we have two variables that are not perfectly correlated, such as height and weight, scores from two achievement tests, or a score from an ability test and an achievement test. Then there exists a tendency for students who make the highest scores on achievement test number one to make less superior scores on test number two. Pupils who make high scores on ability tests tend to make not so high scores on achievement tests or school marks.

What are the implications of this phenomenon for teaching? Obviously we need norms for gains as well as for status and norms based on each initial score. Teachers can develop their own norms, and also teachers can keep records of pupils in order to interpret gains more validly. For example, it may be that a gain of 40 percent more correct items from a low score may be generally expected while a gain of 20 percent more correct from a higher score may represent an exceptional improvement. In comparing gains we must be aware of the initial score. Was it low or high?

One other reason why we should exercise care in the use of gain scores is the error factor. The initial and final test measurements each contain error. Errors then accumulate when the scores are subtracted to determine the gain. The difference score may, therefore, be a representation of error rather than gain. Seldom are tests available to measure reliable short term gains for individual pupils. It is possible, however, to use the difference between means of initial and final testing to determine the effectiveness of instruction for a class.

In summary, a valid test for a teacher's purpose measures what he wants it to measure. He cannot depend on the title for this assurance. He must search further and examine the items in addition to reading the information provided in test manuals. A valid test must also be a reliable test. On the other hand, a reliable test isn't necessarily a valid one because it needs only measure consistently what it does measure. Reliability is a necessary but not a sufficient condition for a good test. However, no test is perfectly reliable. Each contains some error, and it is important that we are cognizant of this fact when making decisions about children.

Gains on test scores must be interpreted with considerable care because of the regression effect. Gains have different meanings for inital low and high scores. It has been suggested that norms be provided that are based on the magnitude of these initial scores.

REFERENCES

1. Davis, F. B. *Identification and Measurement of Reading Skills of High School Students.* Washington, D.C.: Office of Education, U.S. Department of Health, Education, and Welfare, 1967.
2. Dressel, Paul L., and Lewis B. Mayhew. *General Education: Exploration in Evaluation.* Washington: American Council on Education, 1954, 59, 99, 128, 166, 204, 227, 237.
3. Galton, Francis. *Natural Inheritance.* London: Macmillan, 1889, 95.
4. Galton, Francis. *Natural Inheritance.* London: Macmillan, 1889, 106.

Strategies of Measuring Student Understanding of Written Materials

FRANK J. GUSZAK
The University of Texas at Austin

READING texts and manuals both use a variety of words to describe the products of pupils' understanding of written materials (even though such products are usually tagged "reading comprehension skills"). The taxonomic structure developed by Barrett (1967) appears to gather such behaviors into a useful ordering that permits us to discuss comprehension in a meaningful way.

Accepting for communication and hierarchical arrangement purposes the five major types of comprehension as illustrated in the Barrett "Taxonomy of Cognitive and Affective Dimensions of Reading Comprehension," we view the following ordering:

1.0 Literal Comprehension
2.0 Reorganization
3.0 Inferential Comprehension
4.0 Evaluation
5.0 Appreciation

Because, as Barrett notes, "Appreciation involves all the previously cited cognitive dimensions of reading . . ." and goes beyond this, we shall arbitrarily delete this dimension in order that we might clearly attack the more simplified cognitive concerns of literal, reorganization, inferential, and evaluative comprehension. In so doing, we must certainly acknowledge the extremely critical nature of the appreciative or affective dimension.

In thinking about the task of measuring pupils' comprehension development, the author's experiences have forced dichotomous questions that ask the following:

How *do we* measure the various types of comprehension?

How *should* we measure the various types of comprehension?

The tipoff is immediate that significant differences are perceived between the way we measure and the way in which we should measure. Hopefully, the extent and import of the dichotomy will become apparent as we discuss each major comprehension area as defined by Barrett.

Literal Comprehension

How Do We Measure It?

Literal comprehension is measured through the student's skill in recognizing some literal element while reading or recalling such an element after a selection has been read.

Typically, basal reader programs have inserted recognition type questions designed to guide the pupil's understanding of the most pertinent literal understandings in the various stories. Frequently referred to as the "guided reading strategy," this technique places the teacher in the role of the guide who asks the leading questions in advance of a page, story, etc. As the students respond to the task by searching out the element, the teacher has the opportunity to observe those who are succeeding as well as those who are not. Upon the completion of the search, the teacher checks out her observations via oral questioning. The whole task goes something like this:

Teacher: Find the name of the story.
Child: The trip.
Teacher: O.K. Now read the first page and find out when the Parks were taking a trip and where they were going. (Silence as pupils read).
Child: On June 1st.
Teacher: Good, and where were they going?
Child: To the mountains.

The values of such strategies seem evident in that they seem capable of guiding pupils to the prime elements of the story, a skill these pupils must subsequently employ in a variety of reading materials and tasks.

All values can be exploded if

1. the questions don't direct the pupils to the most important elements, i.e., pupils are directed toward insignificant happenings;
2. the directing questions are inappropriate because they direct when such directional aid is superfluous;
3. a few pupils continuously provide all the answers (something inevitable in almost any group).

Recall-type questions suffer from the same kinds of problems. Thus, teachers possess the dangerous possibility of programing children not only to look for insignificant information but also to remember insignificant facts that may be highlighted to the neglect of basic considerations.

How Should We Measure It?

Prior to any oral or written questioning of pupils it is essential to assess the particular reading content as to its 1) most basic concepts and 2) sequence of events (and their relative importance). Such assessment can indicate whether we should ask one question or many more, as well as the specific nature of the most pertinent questions. Nothing is more defeating than to squeeze a multitude of questions out of something of relatively little significance or meaning to the readers or, conversly, to miss many basic points in a content of particular interest or importance. Thus, the first step to better assessment is to know the content, the backgrounds of the discussants, and the interrelationships of the two.

In oral questioning of the recognition-type it is reasonable to *spot different tasks* by saying:

> John, please see why they were taking this trip.
> Mary, find out where they were going and how long they were going to stay.
> Sue, see if you can find things that tell how each of them feel about the trip.
> Bob, I'm hoping that you can find out. . . .

Because the rapidity of assignment completion will invariably throw a monkey wrench in many teacher-directed group tasks, it's useful to *set purposes in advance on the board* and ask the students to *use a marker technique* (such as a paper clip) to mark the specific

elements as they are found in the reading. By so doing, the rapid readers can complete the assignment and go on to something else while the slowest readers have time to finish tasks. A variation of this procedure allows the children to jot down the page, paragraph, and sentence number of certain elements.

For purposes of measuring recall, group response instruments, such as a color wheel, can permit the teacher to find out precisely which students know the answers to specific fact questions. For example, the color wheel is held by each child, and the teacher asks a question which may be answered by one of three colors:

Teacher: The real winner of the game was *Tom, Bill,* or *Joe.* Show *blue* for Tom, *red* for Bill, and *white* for Joe.
Pupils: (The pupils manipulate the color and flash it on signal to the teacher who notes the responses).

Because written questions suffer from the same types of problems as do oral questions, the use of the cloze technique seems especially valuable as a measure of comprehension for any type of reading material. Cloze tests simply involve the restoration by the pupil of deleted words (usually every fifth or tenth word is deleted from a selection of 250 words or more). For a further discussion of this technique the reader is referred to articles by Bormuth (*2*), Culhane (*3*), and Rankin and Culhane (*5*).

Reorganization

How Do We Measure It?

Research and observation by the author suggest that we don't measure reorganization skill often enough. In a study of the questions asked by certain second, fourth, and sixth grade teachers, it was noted that less than 1 percent of the questions were of the reorganization type (*4*).

In all fairness to the teachers, we must suggest that it is difficult to measure reorganization skills via the oral techniques that characterize reading group discussions. Such tasks take time; and when a single student is asked to reorganize a story through a summary or synopsis, there is little left for the other group members to do than to make some additions or corrections.

Silent strategies, such as the following, seem productive of reorganization measurement:

> *Sequence Tasks.* Students are given pictures, sentences, or paragraphs and are asked to order them by their occurrence in the story. Sequence sets can be constructed for various stories. Some are used in basal workbooks where students are asked to do such things as writing the numerical order of specific events.
>
> *Synopsis, Summary Tasks.* When writing skills are developing to a reasonable degree, pupils can go beyond the arrangement or ordering of prepackaged tasks to do their own summarizations.

Reorganization can't be slighted on the excuse that it is unimportant. It is important. Included in the skill is direction toward economy whereby the student can produce precise (short and accurate) reorganizations essential to effective communication.

Inferential Comprehension

How Do We Measure It?

Before allowing the children to turn to the next page, the teacher asks, "Well, what do you think Jack's going to do?" Instantly the response arrives that Jack is going to "swing from the rope." The logic supporting the use of inferential training is certainly sound because we can think and read more ably when we can accurately anticipate what's coming. By constantly anticipating and seeking verifications of our anticipations we can increase both the speed and accuracy of our reading.

Unfortunately though, most teacher-pupil exchanges of the sort suggested do not tap inference but whether the students have

> listened to another reading group encountering the same bit,
> flipped ahead to see the picture, or
> read the next page.

Consequently, much of the envisioned value doesn't materialize, nor will it materialize unless we rigidly hold every child to the same reading selection and page turning pace.

How Should We Measure It?

Stauffer (*6*) essentially dedicates a text to the means for stimulating thinking about reading, with the keystone being inferential thinking. Called the D-R-T-A, or Directed-Reading-Thinking-Activity, Stauffer describes how teachers can guide pupils to plug in their inferential skills to the smallest of clues, beginning with the title of a story or a picture. After making their inferences, the various pupils proceed to test (by reading) the various predictions. Upon verification, the students further predict and set about further verification, etc. This strategy which continues in this cyclic manner is quite different from the illustration at the beginning of this section because the Stauffer technique fosters genuine inference by various pupils as well as legitimate opportunities for verification.

Essential to carrying out some of D-R-T-A strategy as suggested by Stauffer are the following conditions:

—The availability of multiple sets of readers that won't always be previewed for the slower readers via the fastest readers who read them first. (Multiple adoptions will allow for this.)
—The choice by the teacher or group leader of a significant organizer for inference, e.g., a suggestive title or clue.
—The sampling of a wide variety of conjectures so as to increase the investment of all concerned (in the group).
—The accurate verification of the most precise conjecture.

Pupils will learn to sense when to apply convergent conjectures or divergent conjectures. At times, pupils will realize that they totally missed the significant cues that might guide their anticipations. Still, the exercise will refine the processes of anticipation that are capable of making us either strong or weak anticipators—readers, that is. Good readers are good guessers!

Evaluation

How Do We Measure It?

Have you ever heard or used any of these questions?

—Well, how did you like that story (ending, character, etc.)?
—Would you like to be in a situation like that?

—What kind of boy do you think Bill was?
—Which story did you like best in this unit?

If you have heard or used such, you've surely heard the droning replies of "yes" and "no" as well as the various other judgments such as "good," "bad," and other terms.

There's nothing wrong in asking for evaluations if we ask for the supports of the evaluations. All too often, according to the author's research, we fail to plug in the "why" followup question and ask:

—Why did (or did not) you like the story?
—Why would (or would not) you like to be in a situation like that?
—Why do you think Bill was that kind of boy (whatever kind was indicated)?
—Why did you like that story best?

Perhaps you've gotten the message. If so, when someone asks you about something, be sure to tell them *why* it was good, bad, or indifferent.

REFERENCES

1. Barrett, T. "Taxonomy of Cognitive and Affective Dimensions of Reading Comprehension," unpublished paper, University of Wisconsin, 1967.
2. Bormuth, J. "Comparable Cloze and Multiple-Choice Comprehension Test Scores," *Journal of Reading,* 10 (1967), 291-299.
3. Culhane, J. "Cloze Procedures and Comprehension," *Reading Teacher,* 23 (1970), 410-413.
4. Guszak, F. "Teacher Questioning and Reading," *Reading Teacher,* 21 (1967), 227-234.
5. Rankin, E., and J. Culhane. "Comparable Cloze and Multiple-Choice Comprehension Test Scores," *Journal of Reading,* 13 (1969), 193-198.
6. Stauffer, R. *Directing Reading Maturity as a Cognitive Process.* New York: Harper and Row, 1969.

The Development of a Diagnostic Test of Syntactic Meaning Clues in Reading

ALBERT DAVID MARCUS
New York University

THE PRESENT STUDY was undertaken to develop a diagnostic instrument to measure the understanding of literal meaning by intermediate grade students through the use of syntactic clues within written standard English sentences.

Theoretical Rationale

Recently, theoretical developments in the fields of linguistics and reading have joined to revitalize interest in the way meaning is derived from spoken and written language, and both linguists and reading specialists have placed renewed emphasis on the importance of the sentence as the basic meaning-bearing unit within English.

Standardized silent reading tests employed to evaluate a student's reading ability usually have been divided into a vocabulary subtest and a paragraph comprehension subtest. Most paragraph comprehension tests consist of a series of graded reading selections that vary in level of reading difficulty. The difficulty of the selections is controlled by varying the word structure, sentence structure, vocabulary load, and the content or concept load of the selection. As the selections increase in difficulty within the elementary and junior high school level tests, the words become increasingly complex in stucture, the vocabulary load becomes greater, the concept load becomes greater, and the sentences become longer and more complex. Along with these factors, the paragraphs become longer, and their content load is increased. Because of the number of variables within the tests, a student's errors may be due to his lack of knowledge of any one of these factors singly or in combination with the other factors.

These standardized silent reading tests have not attempted to directly measure a student's literal comprehension of sentences, which is the basis of paragraph development. Although the reading grade obtained from the test results can be a valuable tool for evaluating a student's general competence in reading, this grade level score does not indicate weaknesses in specific skills.

Whitehall and Allen have suggested that although written English is closely related to its spoken counterpart, the written form is a separate system of English. Those syntactic features common to spoken English and formal written English that the student already knows should not create reading difficulties for him, but those syntactic features of formal written English that are unfamiliar to the student may create reading comprehension difficulties.

In order to assist reading progress, a teacher should diagnose those elements of syntax within formal written English that a student does not understand and which may hinder his growth in reading comprehension. To meet this need a diagnostic instrument was developed.

Hypothesis

It was hypothesized that the diagnostic instrument would be a valid and reliable measure of students' ability to understand syntactic structures of predication, structures of complementation, structures of modification, and structures of coordination.

Definition of Terms

The following definitions refer to terms used in the study:

1. *Diagnostic test* refers to a test that is used to locate specific areas of weakness or strength within a larger body of information or skills.

2. *Syntactic structures* refers to those devices and patterns within the language according to which words are combined into larger structures for the expression of meaning in sentences. The five signals of English syntactic structures are word order, prosody, function words, inflections, and derivational contrast. The four types of English syntactic structure as classified by Francis (5) follow:

a. *Structures of modification* consist of two immediate constituents, a *head* and a *modifier.*
b. *Structures of predication* consist of two immediate constituents, a *subject* and a *predicate.*
c. *Structures of complementation* consist of two immediate constituents, a *verbal element* and a *complement.*
d. *Structures of coordination* have two or more immediate constituents, which are syntactically equivalent units joined in a structure which functions as a single unit.

Construction of A Test of Sentence Meaning

"A Test of Sentence Meaning" (to be referred to as ATSM) was constructed as a diagnostic instrument for use in determining a student's understanding of syntactic clues to literal meaning within written standard English sentences. It was devised for use with intermediate grade students who had achieved a minimum of a fifth grade level in word recognition skills on Huelsman's Word Discrimination Test.

As a diagnostic instrument ATSM serves two basic purposes: a) it reveals a student's strengths and weaknesses in syntactic knowledge, and b) it presents the teacher with information for planning a program for teaching specific syntactic skills. By indicating those syntactic structures with which a student has difficulty, the test pinpoints specific skills that need to be taught to the students.

ATSM measures the student's ability to handle selected aspects of the four types of syntactic structure which are described as basic by Francis. In this view, all syntactic structures in sentences are manifestations of one or more of these types. Reading a sentence requires recognition of the relationship indicated in one or more of these structural types; and in actual sentences, structures of one type may be included in a structure of another type. For example, a structure of predication (clause) may include one or more structures of modification.

Francis lists five devices by means of which grammatical meaning is indicated: word order, prosody, function words, inflections, and derivational contrast. ATSM employed instances of all these except prosody, which refers to suprasegmental components that appear

in speech—stress, pitch, and juncture. As a silent reading test, this instrument is not appropriate nor does it attempt to measure clues to meaning given by particular speech intonational patterns.

Two systems of grammatical description were used in developing the test. The Francis version of structural grammar was used to isolate the types of syntactic structure which were selected for testing. This approach assumes that part of the comprehending process is the derivation of meaning from patterns of syntactic structures which may appear as part of sentences.

A transformation-generative theory of grammar was used in developing the test items for the specific skills. This system of grammar holds that the actual sentences of discourse are generated from kernel sentences and that certain pairs of grammatical structures can be used to denote the same meaning. Test items utilizing this theory were devised by factoring sentences into their underlying kernels and by comparing transformations with equivalent meanings.

In order to be sure that a student's errors were due to a lack of knowledge of standard written English, the lexical content and internal punctuation of the sentences within the test items were controlled to keep these aspects of language from interfering with the student's ability to deal with syntactic structures of standard written English.

To help insure that the student was familiar with the lexical or dictionary meaning of the words used in the test items, the words used in the test were limited to the most frequently used words as they occurred in the word lists of Dale, Thorndike and Lorge, and Rinsland.

Selection of specific skills. Francis' classification of the grammatical structures of English was used as a basis for organizing the types of grammatical structures included in ATSM. Research in readability and writings by linguists were investigated for clues to the types of grammatical constructions that might cause problems in reading comprehension.

An initial list of twenty-seven types of grammatical structures was complied from those structures that various researchers had suggested as causes of problems in reading comprehension. This list was reduced to seventeen structures that appeared to be representa-

tive of basic English syntactic structures and which were adaptable to a multiple choice question format. The classification of these structures into the categories of structures of modification, structures of predication, structures of complementation, and structures of coordination is presented in Table 1. In addition to these four categories a fifth category was added for combinations of structures. (See Appendix for sample items for each of the seventeen structures of ATSM.)

TABLE 1

CLASSIFICATION OF THE 17 SPECIFIC SYNTACTIC STRUCTURES INCLUDED IN ATSM WITHIN FRANCIS' CATEGORIES OF ENGLISH GRAMMATICAL STRUCTURES

I.	*Structures of Modification:*
	prepositional phrase as noun, verb, or sentence modifier
	complex sentence where relative clause modifies subject
	complex sentence where relative clause modifies object
	complex sentence where relative clause modifies object of preposition
	complex sentence with two relative clauses
II.	*Structures of Predication:*
	passive voice in simple sentence
	passive voice in complex sentence where relative clause contains passive
	recognition of transformations of nominalizations into active verbs
III.	*Structures of Complementation:*
	direct object/indirect object sequence
	direct object/objective complement sequence
	subjective complement embedded as modifier
IV.	*Structures of Coordination:*
	sentence with coordination of phrases
	sentence with coordination of subordinate clauses
	sentence with coordination of independent clauses
	elliptical structures of coordination
V.	*Combinations of Structures:*
	included clauses as modifiers, subjects, or complements
	combinations of structures

Format of test items. The ability to discriminate between sentence structures that had the same or different meanings was used as the underlying principle for developing the format of test items. In arriving at the correct answer or answers to an item a student had to differentiate between the choices that gave a different meaning and those that gave the same meaning, wholly or in part. Four types of multiple-choice items were derived from this principle (see Table 2).

To produce a meaning-oriented test rather than a usage-oriented test, all of the possible choices were grammatical, even the distractors —those possible choices that were incorrect.

In developing test items, specific syntactic structures were measured within the context of sentences. Lexical units from the original sentence of each item were included in the distractors of that item; but changes in word order, inflectional endings, derivational affixes, and/or function words were used to indicate meanings different from the meanings of the original sentence or the kernel sentences of the original sentences.

Format 1 and Format 2 items were both designed to measure the student's knowledge of transformations that gave equivalent meanings. In a Format 1 item the student had to find the transformation that had the same meaning as the given underlined sentence. In Format 2 items the student was given four sentences, three of which were transformations of one another and denoted the same meaning. The fourth sentence utilized the vocabulary of the other three sentences but denoted a different meaning. The student was to determine which sentence had a meaning different from that of the other three sentences.

Format 3 and Format 4 items were designed to measure the student's knowledge of kernel sentences within certain subordinate and coordinate constructions. In a Format 3 item the student was required to find the two sentences that said something true about the given underlined sentence. This format involved the ability to analyze a given structure into its basic kernel sentences. Without a knowledge of the correct kernel sentences a student would not know the correct meaning of the larger structure. Format 4 items required a student to find the two smaller sentences that gave the whole

TABLE 2

THE FOUR TYPES OF FORMAT FOR TEST ITEMS

Format 1

Directions:

Choose the one sentence that has the same meaning as the underlined sentence.

The man gave the boy a puppy.

a. The man gave away the boy's puppy.
b. The man gave a puppy to the boy.
c. The boy gave a puppy to the man.
d. The man gave a puppy away for the boy.

Format 2

Directions:

Three of the four sentences below have the same meaning. Choose the one sentence that has a *different* meaning.

a. Mother gave the baby the bottle.
b. The baby was given the bottle by mother.
c. The baby gave mother the bottle.
d. The bottle was given to the baby by mother.

Format 3

Directions:

The underlined sentence can be made into smaller sentences. Choose *two* sentences that say something true about the underlined sentence.

Mary saw the boy who ate the pie.

a. The boy saw Mary eat the pie.
b. The boy ate the pie.
c. The boy saw Mary.
d. Mary ate the pie.
e. Mary saw the boy.

Format 4

Directions:

Choose the *two* sentences that combine to give the complete meaning of the underlined sentence.

Bob and Don ate the bread and jelly.

a. Bob and Don ate the bread.
b. Bob ate the bread and jelly.
c. Don ate the bread.
d. Bob and Don ate the jelly.
e. Don ate the jelly.

meaning of the original sentence. Again, the student had to find the kernel sentences that gave the meaning that was equivalent to the original sentence.

Determination of number of items. By using the law of chance probability a procedure was developed for evaluating a student's ability on a group of six multiple-choice items. Depending on the number of items answered correctly, a student's knowledge of each skill was ranked good, fair, or poor.

Content Validity

It should be realized that the results of this study depended fundamentally on the appropriateness of the items used and how well each item fit the grammatical classification for which it was written.

To establish content validity for ATSM the test items were submitted to three linguists who independently evaluated each item. The following criteria were used in judging each item: 1) Each item was to be in fact a structure of the type supposedly being tested. 2) All sentences within a test item were to be natural sentences such as might reasonably occur in normal discourse. 3) In test items designed to check the student's knowledge of kernel sentences within larger sentences, the denotative meaning of the "correct answers" was to be in accord with the denotative meaning of the lead sentence and the denotative meaning of each "incorrect answer" was not to be. 4) In test items designed to check the student's knowledge of transformations with equivalent meanings, the transformations were to denote equivalent meanings and the incorrect answers were to denote a different meaning.

The first draft of ATSM contained 102 test items with six items constructed for each of the seventeen grammatical structures. Four additional items, used as samples in the test directions, were also included for evaluation. After the judges had evaluated the items, nineteen were revised in accordance with the recommendations and none were omitted.

A Test of Sentence Meaning

The seventeen specific grammatical structures included in the test were divided into three sections of approximately the same length so that each could be administered as a separate subtest.

In organizing the test all six items measuring a specific syntactic structure were grouped together. The correct answer or answers for each test item were randomly assigned within the group of answers for that test item by using a table of random numbers.

The initial screening device, Huelsman's Word Discrimination Test, and the three parts of ATSM were administered to 487 boys and girls in the fifth, sixth, seventh, and eighth grades from disadvantaged area schools and middle-class area schools. After the exclusion of the students who failed to complete ATSM or who failed to reach the 5.0 reading level on the screening device, data for 421 students were available for analysis.

Reliability

Reliability coefficients using the Kuder-Richardson Formula 20 were computed for all 102 items for each of the four grades and the total test sample. This formula measures the internal consistency of the test, and the intercorrelations of the items is the essential source of this kind of reliability.

The reliability coefficients for ATSM ranged from .95 for grade five to .89 for grade eight (see Table 3). These coefficients were

TABLE 3

RELIABILITY COEFFICIENTS, STANDARD DEVIATIONS, AND STANDARD ERRORS OF ATSM FOR EACH GRADE AND TOTAL TEST SAMPLE

Grade	*Reliability Coefficient*	*Standard Deviation*	*Standard Error*
Grade 5	.95	18.04	5.6
Grade 6	.90	13.00	5.7
Grade 7	.91	12.75	5.3
Grade 8	.89	10.77	4.8
Total Sample	.94	16.22	5.4

considered to represent a satisfactory level of reliability for ATSM, since the reliability coefficients of standardized tests usually range within .80s and .90s.

Grade Averages for Total Test

As is generally expected, students at higher grade levels achieved higher scores on the total test and on each of the seventeen structures of ATSM than students at lower grade levels. The grade averages for the total test increased from the fifth grade to the eighth grade

TABLE 4

AVERAGE NUMBER OF ITEMS CORRECT FOR EACH GRADE*

Grade	*Average Number of Items Correct*
5	60
6	66
7	73
8	81

*ATSM included a total of 102 items.

with the eighth grade students correctly answering about twenty-one more items than the fifth grade students. The mean percent correct increased from the fifth grade to the eighth grade for each of the seventeen structures of ATSM. These results serve to lend further support for the validity of ATSM.

Analyses and Results

Analyses of the students' mistakes indicate that a number of syntactic features were related to the students' inability to derive the appropriate meaning from sentences. During the test administration some of the students regressed to earlier reading habits by vocalizing the test items. This occurrence might further indicate the students' need for intonation clues in the derivation of meaning in written material.

Complex sentences in which a relative clause interrupted the subject-verb-object sequence of the independent clause were more difficult for the students to understand than complex sentences whose basic components were not separated. In deriving the meaning of complicated sentences some students mistakenly thought that a coincidental noun-verb-noun sequence of words was a subject-verb-object sequence and thus a kernel sentence of the larger sentence.

Some students did not distinguish between denotated literal meanings and implied meanings. Other errors indicated also that some students did not understand the semantic and/or the syntactic meaning of various function words, including simple and compound prepositions, correlatives, and relative pronouns.

The test served its diagnostic purpose by indicating those syntactic structures with which an individual student had difficulty. This information enabled the teacher to plan a specific program for those students who needed additional help.

APPENDIX

Sample Items for the Seventeen Grammatical Structures included in ATSM

I. *Structures of Modification*

Directions: Choose two sentences that say something true about the underlined sentences.

Prepositional phrase as noun, verb, or sentence modifier:

Jane gave the cooky behind the jar to the boy.

a. Jane gave the boy the cooky.
b. The giving of the cooky was behind the jar.
c. Jane was behind the jar.
d. The cooky was behind the jar.
e. The boy was behind the jar.

Complex sentence where relative clause modifies subject:

The boy to whom she gave the rabbit climbed through the hole in the fence.

a. The boy climbed through the hole in the fence.
b. The boy gave her the rabbit.
c. The rabbit climbed through the hole in the fence.
d. She gave the rabbit to the boy.
e. She climbed through the hole in the fence.

APPENDIX *(Continued)*

Complex sentence where relative clause modifies object:

The dog frightened the child whom the workman was protecting.

a. The workman was protecting the dog from the child.
b. The child was frightened by the workman.
c. The dog was protecting the child from the workman.
d. The workman was protecting the child.
e. The dog frightened the child.

Complex sentence where relative clause modifies object of preposition:

The uncle of the boys who were swimming in the river drowned in a boat accident yesterday.

a. The uncle drowned in a boat accident yesterday.
b. The boys were swimming yesterday.
c. The boys were swimming in the river.
d. The boys drowned in a boat accident in the river.
e. The boys drowned while swimming in the river.

Complex sentence with two relative clauses:

The woman whom Uncle Robert admired handed the gift to the doctor whom she visited.

a. The woman visited the doctor.
b. Uncle Robert admired the gift.
c. Uncle Robert admired the woman.
d. Uncle Robert handed the gift to the woman he admired.
e. The doctor visited the woman.

II. *Structures of Predication*

Directions: Three of the four sentences below have the same meaning. Choose the one sentence that has a *different* meaning.

Passive voice in a simple sentence:

a. He gave the candy to the lady.
b. He was given the candy by the lady.
c. The lady gave him the candy.
d. The candy was given him by the lady.

Directions: Choose two sentences that say something true about the underlined sentence.

Passive voice in complex sentence where relative clause contains passive:

The men who were attacked by the police ran around the corner.

APPENDIX *(Continued)*

a. The police ran around the corner.
b. The men ran around the corner.
c. The men attacked the police.
d. The police were attacked.
e. The police attacked the men.

Directions: Three of the four sentences below have the same meaning. Choose the one sentence that has a different meaning.

Recognition of transformations of nominalizations into active verbs:

a. Bob's instructions to her were to arrange for the wedding's quick conclusion.
b. Bob instructed her to quickly conclude the wedding arrangements.
c. Bob instructed her to quickly conclude the arrangements for the wedding.
d. Bob's instructions to her were that the wedding arrangements were to be brought to a quick conclusion.

III. *Structures of Complementation*

Directions: Choose the one sentence that has the same meaning as the underlined sentence.

Direct object/indirect object sequence.

He brought the woman her son.

a. He brought the woman with her son.
b. He brought the woman and her son.
c. He brought the woman to her son.
d. He brought her son to the woman.

Directions: Choose two sentences that say something true about the underlined sentence.

Direct object/objective complement sequence:

The committee appointed her brother president.

a. The president appointed her brother.
b. The committee appointed her brother to be president.
c. The committee and the president appointed her brother.
d. The committee appointed her.
e. The committee appointed her brother.

APPENDIX *(Continued)*

Subjective complement embedded as modifier:

The old man outside owns a small cat.

a. The old man owns a small cat.
b. The old man's cat is outside.
c. The old man is outside.
d. The cat that the old man owns is outside.
e. The old man owns the small cat outside.

IV. *Structures of Coordination*

Directions: Choose the two sentences that combine to give the complete meaning of the underlined sentence.

Sentence with coordination of phrases:

Jane and Tom ran and jumped along the road.

a. Jane jumped along the road.
b. Jane ran and jumped along the road.
c. Tom ran along the road.
d. Tom jumped along the road.
e. Tom ran and jumped along the road.

Directions: Choose two sentences that say something true about the underlined sentence.

Sentence with coordination of subordinate clauses:

The horse jumped because he saw the snake and because the rider frightened him.

a. The rider frightened the snake.
b. The horse saw the snake.
c. The snake frightened the rider.
d. The rider frightened the horse.
e. The rider saw the snake.

Compound sentence:

She is not only intelligent, but she is also beautiful.

a. She is not intelligent.
b. She is not intelligent, but she is beautiful.
c. She is beautiful.
d. She is only intelligent.
e. She is intelligent.

APPENDIX *(Continued)*

Elliptical structures of coordination:

Anne asked Jane to come at six and Mary at noon.

a. Anne asked Jane to come at six and at noon.
b. Anne asked Mary at noon.
c. Jane was to be at Mary's at noon.
d. Anne asked Jane to come at six.
e. Anne asked Mary to come at noon.

V. *Combinations of Structures*

Directions: Three of the four sentences below have the same meaning. Choose the one sentence that has a different meaning.

Included clauses as modifiers, subjects, or complements:

a. Everyone knows that he is a liar.
b. That he is a liar everyone knows.
c. He is a liar that everyone knows.
d. Everyone knows he is a liar.

Directions: Choose two sentences that say something true about the underlined sentence.

Combinations of structures:

Mary complained that no one was helping her clear off the tables in the dining room since the group decided that Betty should be relieved of housekeeping duties because she cooked the meals.

a. The group decided to help Mary clear off the tables.
b. Mary cooked the meals.
c. Mary was relieved of her housekeeping duties.
d. Mary complained that no one helped her clear the dining room tables.
e. The group decided to relieve Betty of housekeeping duties.

REFERENCES AND NOTES

1. Allen, Robert L. "Better Reading through the Recognition of Grammatical Relations," *Reading Teacher,* 18 (December 1964), 194-198.
2. Allen, Robert L. "Written English Is a 'Second Language,'" *English Journal,* 55 (September 1966), 739-746.
3. Bach, Emmon. *An Introduction to Transformational Grammars.* New York: Holt, Rinehart and Winston, 1964.

4. Dale, Edgar, and Jeanne S. Chall. "A Formula for Predicting Readability," *Educational Research Bulletin,* 27 (January-February 1948), 11-20, 37-54.
5. Francis, W. Nelson. *The Structure of American English.* New York: Ronald Press, 1958, 425-426.
6. Guilford, J. P. *Fundamental Statistics in Psychology and Education.* New York: McGraw-Hill, 1956.
7. Huelsman, Charles B., Jr. "The Visual Perception of Word Form," unpublished doctoral dissertation, University of Chicago, 1949.
8. Huelsman, Charles B., Jr. *Word Discrimination Test: Form A.* Oxford, Ohio: Miami University Alumni Association, 1958.
9. Rinsland, Henry D. *A Basic Vocabulary of Elementary School Children.* New York: Macmillan, 1945.
10. Thomas, Owen. *Transformational Grammar and the Teacher of English.* New York: Holt, Rinehart and Winston, 1965.
11. Thorndike, Edward L., and Irving Lorge. *The Teacher's Word Book of 30,000 Words.* New York: Bureau of Publications, Teachers College, Columbia University, 1944.
12. Whitehall, Harold. *Structural Essentials of English.* New York: Harcourt, Brace and World, 1956.

What Do Diagnostic Reading Tests Really Diagnose?

CAROL K. WINKLEY
Northern Illinois University

IN RESPONSE to the question "What is a good diagnostic reading test?" the answer frequently given is "There is none!" If the inquirer is seeking a single instrument equally appropriate for all levels and suitable for locating problems in all skill areas, the response is, no doubt, a valid one. Yet, there is a need for instruments that classroom teachers can use to supplement their judgments based on diagnostic teaching and that clinicians can use to pinpoint areas when conducting a clinical diagnosis. What is available?

In order to answer this question, a study was made of nine reading tests, including those planned for both individual and group administration, which are claimed to be chiefly diagnostic instruments. Titles of the tests examined are found at the top of the columns in Table 1. The first entry across the table provides information regarding the grade and/or reading levels for which the test was intended. The second entry indicates whether the test must be given individually or whether it can be used in a group situation. A careful analysis of each subtest and its stated or implied purpose revealed that these nine instruments contained subtests for the following:

1. Measuring potential reading level.
2. Measuring silent and oral reading performance.
3. Estimating independent and instructional reading levels.
4. Identifying inhibiting factors.
5. Determining chief area of skill deficiency.
6. Determining technique of word identification.
7. Locating word recognition difficulties.

The author wishes to acknowledge the assistance of Benita Vyverberg, graduate assistant, in collecting the data for this study.

It is clear that no common definition of a diagnostic reading test is held by the authors of these tests, nor have the various authors had similar purposes in mind as they developed their tests.

Assessments of Potential Reading Level

Six instruments contain subtests purporting to estimate the child's potential level of reading achievement. Three general types of activities were utilized by the various authors:

1. Listening comprehension of paragraphs read aloud by the teacher (D), (Sp).
2. Selecting appropriate meanings of words presented orally (GM), (St I), (St II).
3. Selecting word opposites as words are read aloud (B).

No doubt the subtests described are included in these diagnostic batteries to enable the teacher to easily determine whether each child is disabled in reading (reading at a level significantly below his ability level).

Measures of Silent and Oral Reading

Two instruments, (GM) and (Sp), provide a subtest of Oral Reading while the Durrell Analysis of Reading Difficulty includes a subtest of "Silent Reading" in addition to "Oral Reading." In each instance, successive paragraphs, each increasing in difficulty over the previous one, are read. As a measure of unaided recall, the child retells each story to the examiner in the "Silent Reading" section of Durrell's test. A simple comprehension check follows each paragraph, except in the Gates-McKillop battery.

These subtests appear to have three possible purposes: 1) to provide an opportunity to record and analyze types of oral reading errors; 2) to make it possible to compare difficulties in silent reading with those in oral reading, as in Durrell's test; and 3) to make it possible to compare reading achievement with some measure of ability in order to determine whether a child is truly disabled in reading.

TABLE 1

An Analysis of Subtests of Diagnostic Reading Tests

	Bond, Balow, & Hoyt, *Silent Reading Diagnostic Tests* (BBH)	*Botel Reading Inventory* (B)	*Durrell Analysis of Reading Difficulty* (D)	*Gates-McKillop Reading Diagnostic Tests* (GM)
1. Reading and/or Grade Levels	Reading Levels 2-6	Not Stated-Probably PP-6	Nonreader to 6th Level	Probably Nonreader to 6th
2. Method of Administration	Group	Individual & Group	Individual	Individual
3. Subtests Measuring Potential Reading Level		Word Opposites Listening Test	Listening Comprehension	Oral Vocabulary
4. Subtests Measuring Silent & Oral Reading			Silent Reading Oral Reading	Oral Reading
5. Subtests Estimating Independent & Instructional Reading Levels		Word Recognition Test (PP-4) Word Opposites Test		
6. Subtests Identifying Inhibiting Factors			Visual Memory of Words (primary) Visual Memory of Words (intermediate) Hearing Sounds in Words Learning to Hear Sounds in Words Phonic Spelling of Words Spelling Test	Auditory Discrimination Auditory Blending Spelling
7. Subtests to Determine Area of Skill Deficiency		Word Opposites Tests (reading and listening)		Oral Vocabulary
8. Subtests to Determine Technique of Word Identification			Word Recognition & Word Analysis	Words: Flash Presentation Words: Untimed Presentation Phrases: Flash Presentation
9. Subtests to Locate Word Recognition Difficulties	Words in Isolation Words in Context Visual-Structural Analysis Syllabication Word Synthesis Beginning Sounds Ending Sounds Vowel & Consonant Sounds	Consonant Sounds Consonant Blends Consonant Digraphs Rhyming Words Long & Short Vowels Other Vowel Sounds Number of Syllables Accented Syllables Nonsense Words	Letters Hearing Sounds in Words Sounds of Letters Phonic Spelling of Words	Recognizing & Blending Common Word Parts Giving Letter Sounds Naming Capital Letters Naming Lowercase Letters Nonsense Words Initial Letters Final Letters Vowels Auditory Blending Syllabication

TABLE 1 (CONTINUED)

McCullough Word Analysis Tests (Mc)	*Roswell-Chall Diagnostic Reading Tests* (RC)	Spache, *Diagnostic Reading Scales* (Sp)	*Stanford Diagnostic Reading Test,* Level I (St I)	*Stanford Diagnostic Reading Test,* Level II (St II)
4th Level & above	Levels 2-6	Elementary School Levels	Grades 2, 3, & 4[1]	Grades 4[2], 5, 6, 7, & 8[1]
Group	Individual	Individual	Group	Group
		Reading Passages (listening)	Vocabulary (oral)	Vocabulary (oral)
		Reading Passages (oral)		
		Word Recognition Reading Passages (oral)		
			Auditory Discrimination	
			Vocabulary (oral) Reading Comprehension	Vocabulary (oral) Reading Comprehension a. Literal b. Inferential Rate of Reading
			Word Recognition	
Initial Blends & Digraphs Phonetic Discrimination Matching Letters to Vowel Sounds Sounding Whole Words Interpreting Phonetic Symbols Dividing Words into Syllables Root Words in Affixed Forms	Single Consonants & Combinations Vowel Sounds Rule of Silent "e" Vowel Combinations Syllabication	Consonant Sounds Vowel Sounds Consonant Blends Common Syllables Blending Letter Sounds	Syllabication Beginning & Ending Sounds Blending Sound Discrimination	Syllabication Sound Discrimination Blending

Estimates of Independent and Instructional Levels

Recognizing the necessity for teachers to be able to locate and then provide instruction at the appropriate levels, both Botel and Spache include subtests in their diagnostic batteries to be used to estimate independent and instructional levels. Each author includes graded word lists and specific criteria to be applied in estimating reading levels. Spache also includes graded reading passages for the child to read aloud while Botel incorporates a "Word Opposites Reading Test" into his battery. The remaining authors of diagnostic reading tests make no provision for determining the actual reading levels of the pupils taking their tests.

Identifiers of Inhibiting Factors

Inhibiting factors are those characteristics of the child and/or the home and school environment which are preventing normal progress in reading. Correcting or alleviating them will make it possible for the child to learn to read with greater ease, but the reading difficulty itself must still be identified and skill deficiencies eliminated through remedial teaching.

Both the Durrell Analysis of Reading Difficulty and the Gates-McKillop Reading Diagnosic Tests contain subtests of certain visual and/or auditory aptitudes. The administration of three subtests of the Durrell battery makes it possible to determine whether a child has strengths or weaknesses in both visual and auditory skills. At the primary level the subtest "Visual Memory of Words" provides an evaluation of a child's ability to select a word, seen in a brief exposure using a tachistoscope, from several words of similar configuration. "Hearing Sounds in Words" is a subtest requiring the child to select the word printed in the test booklet that begins, ends, or begins and ends with the same sound(s) heard in words pronounced by the examiner. An analysis of the child's errors on a third subtest ("Spelling Test"), will often disclose additional information concerning the child's relative use of visual memory and phonic principles in writing words having both regular and irregular spelling.

Three similar subtests, which are much more difficult, are

provided for students reading at the intermediate grade levels: 1) "Visual Memory of Words—Intermediate," which requires the child to write the word seen in a brief tachistoscopic exposure; 2) "Phonic Spelling of Words," in which the child is asked to spell words just as they sound (credit is given for any type of phonetic spelling); and 3)"Spelling Test." In order to identify visual and/or auditory strengths and weaknesses, the same types of comparisons can be made as those suggested at the primary level.

In the Gates-McKillop battery, three subtests of auditory skills are found: 1) "Auditory Discrimination"; 2) "Auditory Blending"; and 3) "Spelling." With a sample of fourteen items, the child is asked to tell whether two words pronounced by the examiner are the same or different words. As an indication of a child's auditory blending ability, he is asked to pronounce as a whole a word which he has heard the examiner say part by part. On the spelling test the words are spelled aloud by the child to enable the examiner to determine whether the child spells letter by letter or by phonic elements.

"Auditory Discrimination" is also a subtest of the Stanford Diagnostic Reading Test, Level 1. The format of this test differs from that of the two tests of auditory discrimination mentioned earlier, thus making it possible to administer the test to groups. After the teacher pronounces two words, the child makes an *X* through *B* in his test booklet if the words begin the same, through *E* if they end the same, and through *M* if the middle sounds are the same.

Determiners of Chief Area of Skill Deficiency

In making a diagnosis of a child's reading difficulties, per se, the diagnostician's first task is to determine the chief area of skill deficiency as being in word recognition, vocabulary or word meanings, quality of comprehension, or rate of comprehension. It is estimated that 90 to 95 percent of the children who have trouble with reading have deficiencies in the area of word recognition, deficiencies which in turn affect obtaining the meanings of the words, understanding what is read, or the speed of reading. This statement means that 5 to 10 percent of disabled readers will not have any major problems in the area of word recognition and can be expected to have as their

chief area of weakness either vocabulary, comprehension, or rate. Do diagnostic reading tests help to determine a child's chief area of skill deficiency?

There are subtests in four of the diagnostic instruments examined which will give the diagnostician some help. The remaining test batteries have subtests in one skill area only—word recognition. Each of the four instruments mentioned above includes some measure of vocabulary but only the Stanford Diagnostic Reading Tests include subtests of comprehension and rate.

The Stanford tests and the Gates-McKillop Reading Diagnostic Tests each include a listening test of vocabulary described earlier as a measure of potential reading level. The child is required to do no reading. Thus information regarding his knowledge of word meanings which is not hampered by inability to attack unknown words met in silent reading is provided. If the grade score appears to be low in comparison with the child's performance on various word recognition subtests, there may be evidence that vocabulary should be considered the child's chief problem area.

The "Word Opposites Tests" (Reading and Listening) of the Botel Reading Inventory are not so much measures of comprehension, as the author states in his manual, as they are tests of vocabulary or knowledge of word meanings. He suggests that a comparison of scores obtained when the test is read silently with scores earned when the teacher reads the words aloud would help to identify those pupils whose reading performances were significantly lower than their potentials. However, these tests can serve another purpose. If the listening score is considerably higher than the reading score, the child can be suspected to be weak in word recognition rather than vocabulary.

In the "Reading Comprehension" subtests of the Stanford Diagnostic Reading Tests, numbered blanks appear in paragraphs which the child reads silently, selecting from four choices the word that belongs in each space. Level two contains a subtest, "Rate of Reading," in which children are timed in reading content of uniform difficulty and selecting an appropriate word from three choices in every third line to fit the meaning of the sentence. After the raw scores have been converted to stanines, the stanine ratings can be compared. A difference of two or more stanines between subtests is indicative of a possible area of skill deficiency.

"Diagnosers" of Difficulties in Vocabulary, Comprehension, and Rate

Compared with word recognition, few disabled readers have major difficulties in the areas of vocabulary, comprehension, and rate. Those students who do have trouble in these areas are most frequently found at the junior and senior high levels. Nevertheless, instruments are needed to determine whether a child's vocabulary difficulties are due to a lack of understanding of prefix and suffix meanings, not knowing multiple meanings of words, or lack of dictionary skills. There are none. The same situation exists when one looks for a diagnostic test of "Rate." There is no test to determine a child's flexibility of speed when reading for different purposes, for example.

The only diagnostic reading test to provide a breakdown of the child's comprehension skills was the Stanford Diagnostic Reading Test, Level II. About half of the items test literal comprehension, and the remaining items check inferential comprehension. When one comprehension score is two or more stanines below the other score the child may need remedial instruction in that area of comprehension. The examiner still will not know whether the child needs help in understanding main ideas of selections, in understanding sequence, or in recalling facts or details, for example, even though he has been found weak in literal comprehension. Neither are subskill deficiencies identified in the broad area of inferential comprehension.

Determiners of Technique of Word Identification

All of the diagnostic instruments that were examined contained subtests, listed in Table 1, which assessed a variety of word recognition skills. Not all of the tests included subtests of instantaneous word recognition which could then be compared with another subtest in which the child was given sufficient time to use his phonic and structural analysis skills to attack the words not recognized at sight. Such a comparison makes it possible for the examiner to determine not only the size of each child's sight vocabulary but the extent to which he can use various word recognition skills. If the child's knowledge of phonic and structural analysis skills has not developed

TABLE 2

PRETEST LEVELS OF SUBTESTS ON DIAGNOSTIC READING TESTS

1. Locating grapheme presented visually in written words.
2. Isolating a sound heard in a spoken word.
3. Selecting words from several presented auditorally that contain a certain sound.
4. Naming letters.
   - (D) Letters (consonants & vowels)
   - (GM) Naming Capital Letters (consonants & vowels)
   - (GM) Naming Lowercase Letters (consonants & vowels)
5. Selecting grapheme (from several) representing sound in word pronounced by examiner.
   - (BBH) Beginning Sound (consonant combinations)
   - (BBH) Ending Sounds (vowels)
   - (BBH) Vowel and Consonant Sounds
   - (GM) Initial Letters (consonants)
   - (GM) Final Letters (consonants)
   - (GM) Vowels
   - (Mc) Initial Blends & Digraphs
   - (Mc) Matching Letters to Vowel Sounds
   - (ST I) Blending (consonants, consonant combinations, & vowels)
6. Selecting printed word containing phoneme heard in word.
   - *(BBH) Words in Context
   - (D) Hearing Sounds in Words (consonants & consonant combinations)
   - (GM) Nonsense Words (consonants, consonant combinations, & vowels)
   - (Mc) Phonetic Discrimination (vowels)
7. Selecting grapheme standing for sound in name of a pictured object.
   - (BBH) Words in Isolation (consonants, consonant combinations, & vowels)
   - (ST I) Beginning and Ending Sounds (consonants, consonant combinations, & vowels)

7a. Selecting word containing a different grapheme from that representing a particular phoneme in a written word.
   - *(Mc) Interpreting Phonetic Symbols
   - (ST I) Sound Discrimination
   - (ST II) Blending

8. Writing grapheme representing sound heard in word pronounced by examiner.
   - (B) Consonant Sounds
   - (B) Consonant Blends
   - (B) Consonant Digraphs
   - (B) Long & Short Vowels
   - (B) Other Vowel Sounds

TABLE 2 *(Continued)*

	(D)	Phonic Spelling of Words (consonants, consonant combinations, vowels, & vowel clues)
	*(Sp)	Letter Sounds (consonants & vowels)
9.	Giving sound represented by letter.	
	(D)	Sounds of Letters (consonants & consonant combinations)
	(GM)	Giving Letter Sounds (consonants & vowels)
	(RC)	Single Consonants & Combinations
	(RC)	Vowel Sounds
	(Sp)	Consonant Sounds
	(Sp)	Consonant Blends
10.	Pronouncing unknown words.	
	*(B)	Rhyming Words
	(B)	Nonsense Words (consonants, consonant combinations, vowels, & vowel clues)
	(GM)	Blending Words Parts (consonants, consonant combinations, vowels, & vowel clues)
	*(Mc)	Sounding Whole Words (consonants, consonant combinations, vowels & vowel clues)
	(RC)	Short Vowel Sounds
	(RC)	Rule of Silent "e"
	(RC)	Vowel Combinations
	(Sp)	Vowel Sounds
	(Sp)	Common Syllables (consonants & vowels)
	(ST II)	Blending (consonants & vowels)

*Variations of behavior called for in Pretest Step.

NOTE: Pretest Steps are taken from Winkley, Carol K., "Why Not an Intensive-Gradual Phonic Approach," *Reading Teacher*, 23 (April 1970), 611-617, 620.

to a point where he can use these skills in attacking unknown words, he has not yet acquired them and needs further instruction.

Subtests providing the opportunity to compare flash presentations of words with untimed presentations are found in three diagnostic batteries, (D), (GM), and (Sp). A hand tachistoscope is used in the first two instruments whereas the examiner merely checks words that a child recognizes instantaneously as he reads lists of words in Spache's "Word Recognition" subtest. In each instance the child is given more time to carefully analyze any word not recognized at sight.

Locators of Phonic Problems

All of the diagnostic reading tests analyzed contained three or more subtests of word recognition skills. (See Table 1.) In order to evaluate each test battery as it would function in the identification of a child's chief skill deficiencies in word recognition, two steps were taken. First, the pretest steps listed in Table 2 were used to determine the level of understanding required of a child to perform successfully on any subtest. Second, each subtest of phonic skills was then categorized at the pretest level that most nearly approximated the behavior expected of the testee.

An examination of Table 2 reveals that no subtests were categorized at the lowest three pretest steps. Examples of such behaviors are often called for as a part of a readiness evaluation. Subtests requiring naming of capital and/or lowercase letters are found on the Durrell Analysis of Reading Difficulty and the Gates-McKillop Reading Diagnostic Tests.

Selecting the written representation or grapheme (from a group of four or five letters) corresponding to a sound heard in a word pronounced by the examiner is a common response required on group instruments. However, four subtests of the Gates-McKillop Reading Diagnostic Tests, an individual battery, called for a similar behavior. An interesting variation of this technique of requiring the child to match grapheme to phoneme is found in the "Blending" subtest of the Stanford Diagnostic Reading Test, Level 1. The teacher pronounces a word, such as *trick,* for which the child is to select the appropriate beginning, middle, and ending from two choices for each.

Example:	ⓧ tr	ⓧ i	○ ch
	○ br	○ e	ⓧ ck

Pretest Step 6 is only a slight variation of Step 5 requiring the pupil to select a printed word in which the letter appears, instead of a single grapheme, that stands for a particular phoneme (or phonemes) heard in a word pronounced by the diagnostician. A subtest of this type, "Hearing Sounds in Words," appears in the Durrell Analysis of Reading Difficulty. The "Phonetic Discrimination" subtest of the McCullough Word Analysis Tests differs slightly because

the pupil must identify the stimulus words such as *"blow"*, himself, and then find the word among four choices in which he hears the sound of the underlined letters.

Example: out not horse *old*

The absence of an auditory stimulus increases the difficulty of this exercise. A subtest in the Gates-McKillop battery is somewhat different, also, because the child is directed to select from four nonsense words printed in his test booklet the one pronounced by his teacher. For example, the teacher might say, "spə nĕs'." These spellings appear in the test booklet: *spiness* stinacc spiss squents

At a higher level pretest step, where pictures supplant the auditory stimuli, the child is asked to find the grapheme(s) representing the sound(s) heard in the name of a pictured object. A subtest of the Stanford Diagnostic Reading Test, Level 1, utilizes this technique. The child selects the two- or three- letter combination standing for the sounds heard at the beginning or the end of the word represented by the picture. The subtest entitled "Word in Isolation" of the Silent Reading Diagnostic Tests calls for selecting an entire word to go with a picture. The test differs from the ordinary vocabulary test at the primary level in that the foils are not all real words but represent beginning, ending, middle, or orientation errors that a pupil might make. The key for scoring is coded to enable the teacher to classify the types of incorrect choices made by each child.

Like Pretest Steps 5, 6, and 7, Pretest Step 8 is more closely related to spelling than reading. The child is required to recall the grapheme representing a phoneme heard in a word, a task which is a spelling skill—not a reading skill. Only to the extent that word pronunciation and spelling are related can these tests be considered valid measures of a child's use of phonics in pronouncing unknown words.

In the Botel Reading Inventory, the child writes the grapheme representing the phoneme heard at the beginning, end, or middle of a spoken word. Spache, in his Diagnostic Reading Scales, has the pupil write the letter representing isolated phonemes sounded by the teacher. At the intermediate grade level the Durrell Analysis of Reading Difficulty has a subtest requiring the students to write

phonetically certain words not normally appearing in their vocabularies, such as *carpolite*. Any phonetic spelling is judged correct, even *karpulight*.

Giving the sound represented by a separate letter (Step 9) tests a skill needed in reading, and yet adequate performance on this level does not insure the child's ability to blend the sounds and accurately pronounce an unfamiliar word. Since this is an ability that must be checked individually, only the tests developed for individual administration include subtests requiring this behavior of the testees [(D), (GM),(RC),(Sp) see Table 2].

It makes sense that if a diagnostician wants to find out how well a child uses phonic skills to pronounce an unknown word, he should be given some unknown words to pronounce. How can the diagnostician be sure he has selected unknown words? One way is to use nonsense words like those found in subtests of the Gates-McKillop Reading Diagnostic Tests and the Botel Reading Inventory. Several subtests of the Roswell-Chall Diagnostic Reading Tests include real words that are not normally in the sight vocabulary of a child at the lower levels who is having difficulty with reading. Spache has two subtests requiring the child to pronounce groups of letters: 1) "Vowel Sounds," which has several four-letter words, each containing a different vowel letter, to be pronounced first with the long sound of the vowel and then the short sound, and 2) "Common Syllables," many of which are phonograms to be pronounced in isolation.

Since any technique requiring pupils to respond verbally cannot be incorporated into a group instrument, the authors of two group tests have developed subtests which come close to requiring the same behavior of the children taking the test. McCullough in her group test includes a subtest, "Sounding Whole Words," in which the child must select a word from three unfamiliar groupings of letters by sounding each phonetically. As a fourth option he may put a cross in a blank if no word in the row sounds like a word he knows.

Example: *spayss* trayk smay ——

In the "Blending" subtest of Level II of the Stanford tests, a format similar to that in Level I is used. At this higher level however, the teacher does not pronounce each word, but the child must sound

the elements and blend them together to be sure he has put together a meaningful word.

Locators of Difficulties in Structural Analysis

Compared with the number of subtests found in diagnostic instruments that evaluate various levels of a child's phonic knowledge, there are relatively few tests of structural analysis skills. These have been categorized in Table 3 under 1) Locating Root Word; 2) Syllabication; 3) Blending; and 4) Accent. Each subtest was examined to determine whether an auditory or visual stimulus was presented, and the response required was ascertained as oral or written.

Two subtests involve locating the root word in an affixed word. In both instances the child provides a written response to a visual stimulus appearing in group instruments (BBH) and (Mc). In the recent test developed by Bond, Balow and Hoyt, the child is asked to select the root word, among three choices, from which the word appearing in the first column was made ("Visual - Structural Analysis"). In "Root Words in Affixed Forms" (Mc), the children are directed to circle each prefix and suffix. In some of the words, however, the part remaining when the so-called prefix is circled is not a root word. For example, *mend* is not the root word of *commend*; nor does *invite* have *vite* as its root word. Similar errors are noted in the Bond, Balow, and Hoyt subtest.

Subtests of syllabication skills, found in seven of the nine instruments, were classified on three different stimulus-response levels. A variety of behaviors is expected of children:

1. Circling a number to show the correct number of syllables in each word pronounced by the examiner (B);
2. selecting from three choices the correct syllabic division of a word listed in the first column (BBH);
3. drawing a line to separate the two syllables of a word (Mc);
4. selecting the first syllable of words with one or more syllables (St I),(St II);
5. reading multisyllabic words including compound words, affixed words, and words with inflectional endings (RC); and
6. reading nonsense words of two or more syllables (B), (GM).

TABLE 3

TYPES OF STIMULUS—RESPONSE USED IN SUBTESTS OF STRUCTURAL ANALYSIS SKILLS ON DIAGNOSTIC READING TESTS

	Locating Root Word	*Syllabication*	*Blending of Syllables*	*Accent (Phonic Analysis)*
Auditory Stimulus — Oral Response			(GM) Auditory Blending	
Auditory Stimulus — Written Response		(B) Number of Syllables	(ST I) Blending	(B) Accented Syllable
Visual Stimulus — Written Response	(BBH) Visual-Structural Analysis (Mc) Root Words In Affixed Forms	(BBH) Syllabication (Mc) Dividing Words into Syllables (ST I) Syllabication (ST II) Syllabication	(BBH) Word Synthesis (ST II) Blending	
Visual Stimulus — Oral Response		(B) Nonsense Words (GM) Syllabication (RC) Syllabication	(B) Nonsense Words (GM) Blending Word Parts (Sp) Blending	(B) Nonsense Words

To determine pupils' ability to blend the syllables and pronounce a word as a whole, a similar range of types of activities appeared on the various instruments. In response to an auditory stimulus, where the teacher pronounces the various phonic elements in a word separately, the child is expected to respond by pronouncing the word as a whole. From this lowest stimulus-response level (GM), tests of increasing difficulty and complexity appear on other batteries (see Table 3). At the highest level the child pronounces words showing that he can blend their parts [(GM), (Sp)] and also demonstrates this ability in pronouncing the nonsense words on the Botel Reading Inventory.

Only two subtests in Botel's inventory provide any measure of a child's ability to determine the accented syllable. In the first, the child circles the number that shows which syllable is accented in each word that he hears. In the "Nonsense Words" subtest the examiner can observe the child's ability to place the accent on the correct syllable when pronouncing an unknown word. (Although accent is considered a phonic skill affecting vowel sounds rather than word structure, the subtests appeared to lend themselves to the classification scheme used for structural analysis skills.)

Summary and Conclusions

This careful examination of subtests on nine different diagnostic test batteries revealed the following:

1. These instruments have a variety of purposes, several of which are not truly diagnostic in nature.
2. Most of the instruments cannot be used to determine a child's chief area of skill deficiency. Probably, a survey silent reading test is a better instrument.
3. It is not possible to pinpoint specific problems in the areas of vocabulary, comprehension, or rate with these instruments.
4. Although there are many subtests of word recognition skills, most of them really evaluate spelling ability rather than reading ability.
5. Group-administered tests are limited to silent-type activities,

often requiring the child to listen and select or supply graphemic representations of phonemic elements.

6. No single test, group or individual, assesses all subskills of word recognition from knowledge of consonant sounds to ability to select the accented syllable in an unknown word.
7. Skills required to unlock single syllable words are measured more frequently than those required to attack multisyllabic words.
8. Certain errors exist, particularly in the selection of affixed words.

Before the decision is made to use any part of a diagnostic battery, the examiner should ask himself, "Is this test evaluating an ability not measured better by some other instrument specifically developed to determine intellectual capacity or reading level?" It appears that several authors of diagnostic instruments have attempted to be "all things to all people." Shouldn't a diagnostic reading test be one that diagnoses the reading problem itself? Shouldn't it help the diagnostician to find each child's strengths and weaknesses in reading skill development? Shouldn't a diagnostic instrument provide some indication of the level to which a child's acquisition of a specific skill has progressed? Can we be sure a child knows a phonic skill well enough to use it in reading when he demonstrates the ability to use it in a spelling activity?

There are all questions that must be answered by future authors of diagnostic tests. If teaching strategy is to be determined by a careful analysis of each child's performance, subtests of diagnostic instruments must be constructed to pinpoint the child's difficulties in the reading act itself.

RESEARCH AND REFLECTION

Predicting True Reading Gains after Remedial Tutoring

ANITA B. DAHLKE
Wisconsin State University

Part 1: The Problem, Method, and Conclusions

ONE OF the most crucial concerns in the field of education and in our nation is that of widespread retardation in reading skills among elementary, secondary, and college students. The incidence of reading retardation has been estimated by Marksheffel (*21*) to be from 2,500,000 to 5,000,000 children, so severely retarded in reading that they require immediate specialized help. Although estimates vary, it is probable that more than 10 percent of the children of average intelligence in school are reading so inadequately that their total adjustment is impaired (*25*).

Reports from clinical sources reveal a disproportionate percentage of seriously retarded readers among boys as compared to girls. The range of percentages is from approximately 65 percent boys and 35 percent girls to 90 percent boys and 10 percent girls (*12*). In general, research indicates that girls are better readers than boys in the primary grades but that this difference usually diminishes by sixth grade (*32*).

Results of numerous studies of the relationship of reading achievement and intelligence have led to the conclusion that intelligence is a major factor in reading success at all levels. Researches by Bond (*4*), Bond and Fay (*5*), Monroe (*23*), and Strang (*31*) show that this relationship becomes increasingly more pronounced as populations are sampled at succeedingly higher grade levels. Even though intelligence is related to successful achievement in reading, as it is to all other learning, this fact does not necessarily guarantee reading success for the child with a high IQ. Betts (*2*) concluded that eight out of ten retarded readers have normal or superior intelligence.

Kottmeyer (*17*) states that it is not at all uncommon for bright

pupils to develop reading disability, although most remedial readers will be of dull normal or normal IQ.

Since mental test performance is often considered a good predictor of reading achievement, much research has been done, using the Wechsler Intelligence Scale for Children IQ scores and/or subtest patterns, in the area of identification of successful and unsuccessful readers. Results of these studies have been largely inconclusive.

The use of intelligence tests for prediction has been challenged by Harrington and Durrell (*11*), since reading difficulties occur among children at virtually all intellectual levels. Consideration must also be given to the question of whether intelligence tests measure the important perceptual aspects of reading success and failure. In addition, IQ scores of retarded readers are often spuriously low when measured by a group intelligence test which requires reading.

Although research has been done in the area of early identification of children who are, or are likely to become, retarded readers, this identification process needs to be carried one step further to predict which retarded readers will have the greatest potential for growth. With the vast school enrollment and the shortage of trained reading personnel, it is an economic necessity to gear remedial instruction to the growth potential of the students.

There is a dearth of research studies, particularly those measuring either true or residual gains rather than crude gains, which have attempted to predict reading improvement of retarded readers after remedial tutoring.

No studies were found which used true gains as a measure of reading improvement, as does this study. This is a technique, described by Lord (*19*), which is appropriate when a given individual actually is a member of some natural group under consideration, such as the retarded readers in this study. Knowledge that an individual belongs to a certain group constitutes genuine information about that individual. Lord feels that an efficient method of estimation can and should make use of this information.

Statement of the problem

It was the purpose of this study to attempt to predict true reading gains made by retarded readers after remedial tutoring, through the use of selected student variables.

Independent variables utilized included IQ and subtest scores obtained on the Wechsler Intelligence Scale for Children (*35*), pre-tutoring reading levels on the individually administered *Diagnostic Reading Scales* test (*29*), age, sex, grade placement, and parental socioeconomic status.

Definitions

The "retarded reader," as used in this study, is an individual who is retarded in a number of reading skills by one year or more, if in the primary grades, or by two years or more, if older (*30*).

"True reading gains" are distinct from observed gains made between pre- and post-tutoring reading test scores in that a multiple regression equation is used to overcome chance errors of measurement and spurious gains. Thus the "regression to the mean" phenomenon, observed with raw scores, does not occur (*19*).

Design of the Study

The data needed to attain the purpose of this study were obtained from the files of retarded readers who had been referred to the University of Florida Reading Clinic for a diagnostic work-up and tutoring during the years 1954-1967.

Sample

Sixty-two white subjects, fifty-two boys and ten girls, were included in the sample. They ranged in age from six years and nine months, to 15 years and eight months and were in grades one through eight. Any student having severe visual or auditory impairments was automatically excluded from the study, as were those classified as borderline or mentally defective on the Wechsler Intelligence Scale for Children.

Each subject was tutored approximately twenty to twenty-five hours by an experienced tutor during the University of Florida summer reading clinic program or by a clinic staff member for an equivalent number of hours.

Instruments

The test used to measure intelligence was the Wechsler Intelligence Scale for Children (*35*).

Results of the individually administered Diagnostic Reading Scales (*29*) were used as the pre- and post-tutoring reading achievement scores. Sufficient selections were available in the test booklet so that unfamiliar material was used on the post-tutoring test.

Grade placement scores are given on the following three levels: Instructional (Oral Reading); Independent (Silent Reading); and Potential (Auditory Comprehension) (*29*).

> The term "Instructional Level" is used to designate the child's grade level in oral reading. It implies the level and quality of reading which most teachers would find acceptable in group or classroom practice, and the grade level of basal or other reading materials to which the child should or would be exposed in the typical classroom.
>
> "Independent Level" is that grade level of supplementary instructional and recreational reading materials which the pupil can read to himself with adequate comprehension, even though he may experience some word-recognition difficulties.
>
> The "Potential Level" indicates whether a child is capable of understanding materials of even greater difficulty than those he can read orally or silently. This might be considered the level to which his reading can grow under favorable conditions. Theoretically, a pupil can progress to his Potential Level when his difficulties with mechanics or vocabulary are overcome.

Dependent variable

True reading gains at the instructional level were considered the dependent variable.

The true reading gain was calculated for each subject, based on his observed gain between pre- and post-tutoring instructional level scores on the Diagnostic Reading Scales. An ordinary multiple regression equation was used to overcome the chance errors of measurement and spurious correlation existing between initial status on the pretest and gain between pretest and post-test (*19*).

Independent variables

Student characteristics considered in this study as possible predictors of true reading gain follow:

Wechsler Intelligence Scale for Children:

Full Scale IQ (FS-IQ)
Verbal IQ (V-IQ)
Performance IQ (P-IQ)
Subtests: General Information (Info.)
General Comprehension (Comp.)
Arithmetic (Arith.)
Similarities (Sim.)
Vocabulary (Voc.)
Digit Span. (D.S.)
Picture Completion (P.C.)
Picture Arrangement (P.A.)
Block Design (B.D.)
Object Assembly (O.A.)
Coding (Cod.)

Diagnostic Reading Scales:

Independent Level (Ind.)
Potential Level (Pot.)
Difference between Independent and Instructional Levels (Ind.-I)
Difference between Potential and Instructional Levels (Pot.-I)
Difference between Grade placement and Instructional Level (Gr.-I)

Other:

Age
Sex
Socioeconomic Status (Warner Scale Values)

Hypotheses:

I The distribution of subjects will fall equally into three groups:
$V = P$ (within 12.5 points)
$P > V$
$V > P$

II There will be no significant difference in true gain means in reading improvement in the three groups of $V = P$, $P > V$, and $V > P$.

III There will be no significant difference in true gain means in reading improvement among the WISC Full Scale IQ classifications of the subjects.

120-129—Superior
110-119—Bright Normal
90-109—Average
80- 89—Dull Normal

IV There will be no significant predictors of true reading gain from among the twenty-two student characteristics considered.

Summary

In summary, analysis of the data in this study has indicated:

1. Retarded readers are not distributed equally among V = P (within 12.5 points), P > V, and V > P groups.

2. There is no significant difference in true gain means among the three groups of subjects in V = P, P > V, and V > P.

3. There is no significant difference in true gain means among the four groups of subjects classified by WISC FS-IQ as having superior, bright normal, average, or dull normal intelligence.

4. There are five significant predictors of true reading gain, including grade, age, independent reading level, potential reading level, and the difference between independent and instructional levels.

Implications for education

Assuming that definitions and conditions are similar to those in this study, the following implications might be made:

1. The student who appears to make the best true gain in reading after remedial tutoring is the older one, above primary grades, who has developed his Independent reading level well above his instructional level and whose potential level is also above his instructional level.

2. True gains should be considered in measuring reading improvement rather than crude gains, which may mistakenly lead one to believe that the most retarded readers make the best gains due to the "regression to the mean" phenomenon. The use of a multiple regression equation in computing true gains overcomes chance errors of measurement and spurious correlation existing between initial status on the pretest and post-test.

3. As long as retarded readers are of dull normal intelligence,

or above, their intelligence classification should not preclude nor restrict their reading improvement after tutoring.

Implications for research

1. More studies need to be done using true reading gains as the dependent variable, rather than crude score gains. This information would help to clarify the role of "degree of retardation" as a predictor of reading improvement.

2. This study could be replicated using large groups, matched as to number of boys and girls and/or as to V = P, P > V, V > P groups.

3. The true reading gainability prediction equation should be tested on other retarded readers.

TABLE 1

MEAN, RANGE, AND STANDARD DEVIATION OF VARIABLES

Variables	*Mean*	*Range*	*Standard Deviation*
Grade	4.7	1-8	1.89
Age	11.0	6-9-15-8	2.16
V-IQ	100	80-131	11.44
P-IQ	103	76-125	11.47
FS-IQ	102	81-121	10.67
Info.	10	4-19	2.75
Comp.	11	5-19	2.82
Arith.	9	4-17	2.67
Sim.	10	5-17	3.39
Voc.	8	5-16	4.58
D. S.	6	4-13	4.55
P. C.	11.5	5-17	2.97
P. A.	9	7-17	4.93
B. D.	10.5	5-17	2.62
O. A.	8.5	5-15	4.14
Cod.	8.5	5-17	3.12
Ind.	3.5	1.0-6.5	1.84
Pot.	5.8	2.8-8.5	1.53
Gr.-I	2.4	.9-5.4	1.18
Pot.-I	2.9	.7-5.2	1.10
Ind.-I	.7	0-4.0	.82
True Gain	1.0	.3-2.4	.48

Part 2: The Findings

Information regarding the student characteristics considered in this study, except for sex and socioeconomic status, is presented in Table 1.

Verbal IQ *and Performance* IQ

Distribution of subjects. Due to the conflicting results of research investigating the V-IQ and P-IQ characteristics of retarded readers, it was hypothesized that the subjects in this study would be distributed equally among three groups: V = P (within 12.5 points); P > V; and V > P.

The observed and expected frequency distribution of the subjects is shown in Table 2. The null hypothesis was tested by the chi-square method.

TABLE 2

OBSERVED FREQUENCY AND EXPECTED FREQUENCY IN THREE CATEGORIES ASSUMING A UNIFORM DISTRIBUTION IN THE POPULATION

	V = P	P > V	V > P	*Sum*
Observed f	40	16	6	62
Expected f	(20.66)	(20.66)	(20.66)	62
	$X^2 = 18.1 + 1.1 + 10.4$			
	$= 29.6$			
	$P = <.01, 2\ df$	$X^2 = 9.210$		

Since the value of the test statistic falls in the rejection region, the null hypothesis of equal distribution must be rejected.

The WISC standardization data show that two-thirds of the sample had verbal and performance IQs within the range of one standard deviation of 12.5 points. This is also true in the present study. If the 40 subjects who are in the V = P category are disregarded and the distribution of the remaining subjects is tested for equal probability in the P > V and V > P categories, the null hypothesis must again be rejected, since the test statistic falls in the rejection region.

$X^2 = 11.5$
$p = <.01, 1\ df \qquad X^2 = 6.635$

We can, therefore, assume that there is a significant difference in observed frequency between the two remaining categories, a significantly greater number being in the P > V group. This proportion agrees with the results of many studies which indicate that retarded readers tend to have significantly higher P-IQs than V-IQs (*22*), (*6*), (*24*), (*9*), and (*26*).

Consideration must be given to the fact, however, that the child's verbal score may be limited by the same factors which limit his reading performance (*37*) and that he may be compensating for failure in a verbal field by giving greater attention to self-development in the nonverbal area (*30*).

With only sixteen and six subjects in the P > V and V > P groups, respectively, results of this study have contributed little to research conclusions in this area.

Group differences in true gain means. An analysis of variance was made on the true gain means of the three groups, V = P, P > V, and V > P, in order to determine whether there was a significant difference among them. The appropriate data are found in Table 3 and Table 4.

TABLE 3

SIZE, TRUE GAIN MEAN, AND STANDARD DEVIATION OF V-P GROUPS

Group	*Size*	*Mean*	*Standard Deviation*
V = P	40	1.1	.47
P > V	16	.9	.53
V > P	6	.8	.39

Since the F value does not exceed the critical value of $F.05 = 3.15$, we must retain the null hypothesis and conclude that there is no significant difference among the true means of the groups considered.

With only six subjects in the V > P group in this study, definite

TABLE 4

ANOVA TABLE FOR COMPARISON OF TRUE GAIN MEANS AMONG V-P GROUPS

Source of Variation	*SS*	*df*	*MS*	*F*[x]
Between Groups	.46	2	.23	1.01
Within Groups	13.53	59	.23	
Total	13.99	61		

[x]No significant difference at .05 level.

conclusions should not be drawn. The acceptance of the null hypothesis, however, does cast doubt upon the verbal/performance relationship of retarded readers. If it is true that verbal IQ is a better predictor of reading achievement, as claimed by Hage and Stroud (*10*) and Barratt and Baumgarten (*1*), the V > P group should have improved more than the other two groups. On the other hand, Wilson (*37*) feels that the performance IQ is a better indicator of reading potential for retarded readers than the verbal IQ. Therefore, the P > V group should have made the best gains of the three groups.

Results from this study and others (*20*), (*15*), (*32*), indicate that research in the area of V-IQ and P-IQ characteristics of retarded readers is largely inconclusive.

Full-Scale IQ *Groups*

Research tells us that retarded readers are found at all levels of intelligence; therefore, it should be helpful to determine if there is a significant difference in reading gains made at the various levels.

In an attempt to do so, the subjects in this study were grouped according to the Wechsler Full-Scale Classifications, as listed in Table 5. Results of an analysis of variance to test the null hypothesis of equal true gain means among the groups are given in Table 6.

The calculated F value is far less than the critical value of 2.77 at the .05 level of significance; therefore, the null hypothesis must be retained. It must be concluded that there is no significant difference in true gain means among the four FS-IQ groups considered in this study.

It would appear from results that as long as retarded readers are

TABLE 5

SIZE, TRUE GAIN MEAN, AND STANDARD DEVIATION OF FS-IQ GROUPS

Groups	*Size*	*Mean*	*Standard Deviation*
Superior	3	1.2	.40
Bright Normal	16	.9	.37
Average	33	1.1	.53
Dull Normal	10	1.0	.48

TABLE 6

ANOVA TABLE FOR COMPARISON OF TRUE GAIN MEANS AMONG FS-IQ GROUPS

Source of Variation	*SS*	*df*	*MS*	*F*[x]
Between Groups	.49	3	.16	.7036
Within Groups	13.50	58	.23	
Total	13.99	61		

[x]No significant difference at .05 level.

of dull normal intelligence, or above, their intelligence classification should not preclude nor restrict their reading improvement after tutoring.

Correlation matrix of student characteristics

The null hypothesis that there would be no significant predictors of true reading gain was tested by research data computations made at the University of Florida Computer Center.

A stepwise regression program was selected because it would give the desired correlation matrix of student characteristics considered in this study and a multiple linear regression analysis and summary table (Table 11).

In examining the correlation matrix, four independent variables are found to be positively and significantly correlated with true reading gain at the .01 level of significance. They are grade ($r = .75$), age ($r = .68$), independent reading level ($r = .83$), and potential reading level ($r = .66$). The difference between independent and

instructional reading levels (Ind.-I.) is significantly correlated with true reading gain at the .05 level of significance (r = .26).

The null hypothesis must, therefore, be rejected.

Further examination of the matrix indicates the high intercorrelation among these five variables and true gain. Table 7 shows these significant relationships.

TABLE 7

CORRELATION MATRIX OF FIVE SIGNIFICANT VARIABLES AND TRUE GAIN

	Grade	*Age*	*Ind.*	*Pot.*	*Ind.-I*	*True Gain*
Grade	1.00	.96	.77	.73	.42	.75
Age		1.00	.74	.76	.46	.68
Ind.			1.00	.73	.71	.83
Pot.				1.00	.42	.66
Ind.-I					1.00	.26
True Gain						1.00

The variables of grade and age can almost be considered synonymous (r = .96). Since the instructional, independent, and potential reading levels are measured in grade levels, many of the same factors as in grade and age play a significant role in their high intercorrelation.

The highly significant correlation of grade and age with true reading gain has important implications for education. Results of this study indicate that one should concentrate clinical remedial reading efforts on children beyond the primary grades. Kottmeyer (*17*) concurs in this opinion. In the primary grades, children are in the "learning to read" stage and are gradually building a sight vocabulary, learning word attack skills, and building concepts. Since most classwork is oral, the independent level (silent reading) is often no higher than the instructional level (oral reading). Therefore, reading readiness or beginning reading activities should be geared to strengthening any areas of weakness rather than for remediation. In the grades above primary level, the "reading to learn" stage of development, the emphasis changes from oral reading to silent read-

ing. The student now needs to learn to read and to analyze words independently and, hopefully, strives to develop this skill. His experiential and conceptual backgrounds are widening, as are his sight and meaningful vocabularies. He is used to going to school and should have developed a longer span of attention than the primary child. Add to this development the maturity to see the need for becoming a good reader in studying in the content areas, and evidence mounts as to why he is a better candidate for remedial reading instruction than the very young child, should difficulty be encountered. He may have some word recognition difficulties but is able to get enough clues for good comprehension.

Age, grade and sex differences. With grade level having such a high correlation (r = .75) with true reading gain, it was of interest to compute true mean gains and to consider sex differences at each grade level (see Table 8).

TABLE 8

FREQUENCY AND TRUE READING GAIN MEANS OF SUBJECTS GROUPED ACCORDING TO GRADE LEVEL AND SEX

	Number			*True Reading Gain Means (year)*		
Grade	*Boys*	*Girls*	*Total*	*Boys*	*Girls*	*Combined Scores*
1	2	0	2	.3	—	.3
2	6	3	9	.5	.5	.5
3	5	3	8	.8	.9	.9
4	7	0	7	.8	—	.9
5	12	1	13	1.2	1.0	1.2
6	10	3	13	1.0	1.3	1.1
7	6	0	6	1.5	—	1.5
8	4	0	4	2.0	—	2.0
Total	52	10	62	1.03[x]	.91[x]	1.0

[x]No significant difference at .05 level.

Sex was not found to correlate significantly with true reading gains in the matrix. This result may be partly due to the small number of girls in the sample. In order to determine whether there

was a significant difference between the true reading means of boys (1.03) and girls (.91), as given in Table 8, the student's *t*-test was administered. No significant difference between means was found at the .05 level. Similar results were obtained in the studies by Bluestein (*3*), Sinks and Powell (*28*), and Holowinsky (*14*).

Results of an analysis of variance of true gain among the eight grade levels is shown in Table 9.

TABLE 9

ANOVA TABLE FOR COMPARISON OF TRUE GAIN MEANS AMONG GRADE LEVELS

Source of Variation	*SS*	*df*	*MS*	*F*
Between Groups	9.19	7	1.31	14.56[a]
Within Groups	4.80	54	.09	
Total	13.99	61		

[a]Significant at .01 level.

It can be concluded that there is a significant difference at the .01 level of significance, among true reading gains at the grade levels considered in this study.

WISC *scores.* Rather surprisingly, none of the WISC scale or subtest scores showed a significant correlation with true reading gain. In fact, all of them either negatively correlated with true reading gain or had very small positive correlations. This outcome is in contrast to Krippner's study (*18*) in which both the WISC FS-IQ and V-IQ were significantly correlated with reading gain. Bluestein (*3*), also found IQ to be significant in his study. Both of these studies used raw scores gains in reading improvement which may have resulted in the contradictory findings.

Reading levels. As noted previously, the independent reading level (silent reading) and the potential level (auditory comprehension) are significantly correlated with true reading gain at the .01 level, with ($r = .83$) and ($r = .66$) respectively.

The difference between independent and instructional levels (Ind.-I.) is significantly correlated with true reading gain at the .05 level of significance.

Interestingly, the degree of retardation, measured by the difference between grade placement and instructional level (Gr.-I.), was not significantly correlated with true reading gain in this study. Other studies (*3*), have found it a significant predictor of raw score reading gains. The phenomenon of regression to the mean on post-test scores has led many to believe that remediation efforts should be spent on those most retarded in reading since they apparently make the greatest gain after tutoring. Results of this study indicate that this belief is not true. The unique feature of this research is that true reading gains are used, rather than crude score gains. In the process of computing true gains, the multiple regression equation used overcomes chance errors of measurement and spurious correlations existing between initial status on the pretest and gain between pretest and post-test. Therefore, the apparent "regression to the mean" phenomenon never occurs, and there is strong evidence that it is really just a statistical artifact.

Summary table of stepwise multiple regression analysis

In addition to the correlation matrix already discussed, the program selected computes a sequence of multiple linear regression equations in a stepwise manner. At each step, one variable is asked to the regression equation. The variable added is the one which makes the greatest reduction in the error sum of squares and which, if it were added, would have the highest F value.

It is apparent from this summary table that the two most important factors accounting for variance in the true reading gain are the independent reading level and the difference between independent and instructional levels (Ind.-I.), accounting for 68 percent and 21 percent, respectively. The potential level contributes an additional 7 percent to the variance. All three variables have F values significant at the .01 level.

Although the difference between potential and instructional levels (Pot.-I), P-IQ, and FS-IQ have significant F values at the .01 level also, it can be seen that their role in accounting for true gain variance is very minute, as is that of all remaining variables. Surprisingly, V-IQ contributes nothing at all to the true reading gain variance.

TABLE 10

SUMMARY TABLE OF STEPWISE MULTIPLE REGRESSION ANALYSIS

Step Number	*Variable Entered*	*Multiple*		*Increase in* r^2	*F Value*
		r	r^2		
1	Ind.	0.8258	0.6819	0.6819	128.6302x
2	Ind.-I.	0.9448	0.8926	0.2107	115.7195x
3	Pot.-I.	0.9501	0.9027	0.0102	6.0539x
4	Pot.	0.9884	0.9770	0.0743	184.3102x
5	P-IQ	0.9893	0.9788	0.0018	4.6860x
6	FS-IQ	0.9899	0.9800	0.0012	3.2768x
7	Sim.	0.9901	0.9803	0.0003	0.7878
8	P.C.	0.9902	0.9806	0.0003	0.7447
9	Voc.	0.9904	0.9808	0.0003	0.7599
10	P.A.	0.9906	0.9812	0.0004	1.0676
11	Age	0.9907	0.9815	0.0003	0.8757
12	Gr.-I.	0.9910	0.9821	0.0005	1.4809
13	Info.	0.9911	0.9823	0.0002	0.6432
14	Sex	0.9912	0.9825	0.0002	0.5347
15	B.D.	0.9913	0.9827	0.0002	0.4371
16	Arith.	0.9914	0.9828	0.0001	0.2582
17	Grade	0.9914	0.9829	0.0001	0.2279
18	D.S.	0.9914	0.9829	0.0001	0.1808
19	O.A.	0.9915	0.9830	0.0000	0.1044
20	Cod.	0.9915	0.9830	0.0000	0.0474
21	Comp.	0.9915	0.9830	0.0000	0.0317
22	V-IQ	0.9915	0.9830	0.0000	0.0033

xSignificant at the .01 level of significance.

Using the two variables which contribute 89 percent of the variance, independent level and independent-instructional, it is possible to derive a predictive equation from the computer data which might be used to predict approximate true reading gains of readers as defined in this study under similar tutoring conditions, that this equation would be appropriate only for use with retarded readers as defined in this study under similar tutoring conditions. with reading levels and true gain being determined from Diagnostic Reading Scales results.

True gainability prediction equation:

$$\hat{G} = .08 + .33 \text{ (Ind.)} - .38 \text{ (Ind.-I.)}$$

Results of this prediction equation for true gainability in reading have been found to be reasonably accurate when compared to the actual true reading gains of the subjects in this study. In most cases, the predicted and actual true reading gain vary within a range of one month.

Additional findings

Socioeconomic status, as determined by parental occupation classification on the Warner Revised Scale of Occupations (*34*), originally had been intended as an independent variable in this study. Unfortunately, files of eight of the subjects were incomplete and parental occupations were consequently unknown. It was difficult to assign reliable classifications in other cases because the self-described positions were often of vague nature. With incomplete data such as this, it was deemed advisable to leave this variable out of the matrix.

In order to investigate any differences in true means among the subjects from various occupational groups, as classified, a frequency distribution table was compiled (Table 11) and an analysis of variance performed (Table 12).

It must, therefore, be concluded that there is no significant difference among true reading gains means of children of different socioeconomic status considered in this study. Concurring in this

TABLE 11

FREQUENCY AND TRUE READING GAIN MEANS OF SUBJECTS GROUPED ACCORDING TO PARENTAL OCCUPATIONS (WARNER SCALE)

	Prof.	*Prop. & Mgrs.*	*Bus. Men*	*Clerks, Kindred*	*Man. Workers*	*Prot. & Serv.*	*Farmers*	*Un-Known*
Frequency	16	7	3	9	13	0	6	8
True gain Mean	1.0	.9	1.0	1.3	.9	0	1.2	1.0

TABLE 12

ANOVA TABLE OF TRUE READING GAIN MEANS OF SUBJECTS GROUPED ACCORDING TO PARENTAL OCCUPATION (WARNER SCALE)

Source of Variation	SS	df	MS	F[a]
Between Groups	1.14	6	.19	.81
Within Groups	12.85	55	.234	
Total	13.99	61		

[a]Not significant at the .05 level.

opinion are Keshian (*16*), Reid (*26*), and Dukes (*7*). Quite the opposite results were obtained by Sheldon and Carrillo (*27*), Wilson (*36*), and Hill and Giammateo (*13*). It must be remembered that schools reflect the socioeconomic level which they represent. This level includes the learning environment of school facilities, curriculum, materials, teacher preparation, and teacher effectiveness. Unless this environment were matched for subjects from different socioeconomic groups, no realistic comparisons could be made. Even then, the experiential backgrounds of the children would vary so widely that results would still be inconclusive.

Results of this study have contributed little to forming a conclusion in the area of socioeconomic influence on reading improvement, for the parental occupations were unknown for one-eighth of the subjects while many other occupations were vaguely described.

Summary

In summary, analysis of the data in this study has indicated:

1. Retarded readers are not distributed equally among V = P (within 12.5 points), P > V, and V > P groups.
2. There is no significant difference in true gain means among the three groups of subjects in V = P, P > V, and V > P.
3. There is no significant difference in true gain means among the four groups of subjects classified by WISC FS-IQ as having superior, bright normal, average, or dull normal intelligence.
4. There are five significant predictors of true reading gain, including grade, age, independent reading level, potential reading

level, and the difference between independent and instructional levels.

Implications for education

Assuming that definitions and conditions are similar to those in this study, the following implications might be made:

1. The student who appears to make the best true gain in reading after remedial tutoring is the older one, above primary grades, who has developed his independent reading level well above his instructional level and whose potential level is also above his instructional level.

2. True gains should be considered in measuring reading improvement rather than crude gains, which may mistakenly lead one to believe that the most retarded readers make the best gains due to the "regression to the mean" phenomenon. The use of a multiple regression equation in computing true gains overcomes chance errors of measurement and spurious correlation existing between initial status on the pretest and post-test.

3. As long as retarded readers are of dull normal intelligence, or above, their intelligence classification should not preclude nor restrict their reading improvement after tutoring.

Implications for research

1. More studies need to be done using true reading gains as the dependent variable, rather than crude score gains. This information would help to clarify the role of "degree of retardation" as a predictor of reading improvement.

2. This study could be replicated using larger groups, matched as to number of boys and girls and/or to $V = P$, $P > V$, $V > P$ groups.

3. The true reading gainability prediction equation should be tested on other retarded readers.

REFERENCES

1. Barratt, E. S., and Doris L. Baumgarten. "The Relationship of the WISC and Stanford-Binet to School Achievement," *Journal of Consulting Psychology*, 21 (1957), 144.

2. Betts, Emmett A. "Are Retarded Readers Dumb?" *Education*, 37 (1956), 568-575.
3. Bluestein, Venus. "Factors Related to and Predictive of Improvement in Reading and Long-Term Effectiveness of Remediation," *Dissertation Abstracts*, 27 (1966), 1700A.
4. Bond, E. *Reading and Ninth-Grade Achievement*. New York: Bureau of Publications, Teachers College, Columbia University, 1938.
5. Bond, E., and L. C. Fay. "A Comparison of the Performance of Good and Poor Readers on the Individual Items of the Stanford-Binet Scale, Forms L and M," *Journal of Educational Research*, 43 (1950), 457-459.
6. Coleman, James C., and Beatrice Rasof. "Intellectual Factors in Learning Disorders," *Perceptual and Motor Skills*, 16 (February 1963), 139-152.
7. Dukes, Ben Marshall. "Anxiety, Self Concept, Reading Achievement, and Creative Thinking in Four Socioeconomic Status Levels," *Dissertation Abstracts*, 25 (1964), 7076.
8. Feldt, Leonard, and Richard Gunderson. "The Relationship of Differences between V and Non-V Intelligence Scores to Achievement," *Journal of Educational Psychology*, 51 (1960), 115-121.
9. Frommelt, Leo Alois. "An Analysis of the WISC Profiles of Successful and Unsuccessful Readers in the Elementary School," *Dissertation Abstracts*, 25 (1964), 2849-2850.
10. Hage, Dean S., and James B. Stroud. "Reading Proficiency and Intelligence Scores—Verbal and Nonverbal," *Journal of Educational Research*, 52 (1959), 258-261.
11. Harrington, Sister Mary James, and Donald D. Durrell. "Mental Maturity vs. Perception Abilities in Primary Reading," *Journal of Educational Psychology*, 46 (1955), 375-380.
12. Heilman, Arthur W. *Principles and Practices of Teaching Reading*. Columbus, Ohio: Charles E. Merrill, 1967, 407.
13. Hill, Edwin H., and Michael C. Giammateo. "Socioeconomic Status and Its Relationship to School Achievement in the Elementary School," *Elementary English*, 40 (1963), 265-270.
14. Holowinsky, Ivan. "The Relationship between Intelligence (30-110 IQ) and Achievement in Basic Educational Skills," *Training School Bulletin*, 58 (1961), 14-21.
15. Hopkins, Kenneth D., and William B. Michael. "The Diagnostic Use of WISC Subtest Patterns," *California Journal of Educational Research*, 12 (1961), 116-117.
16. Keshian, Jerry G. "How Many Children Are Successful Readers?" *Elementary English*, 38 (1961), 408-410.
17. Kottmeyer, William. *Teacher's Guide for Remedial Reading*. Webster Publishing, 1959.

18. Krippner, Stanley. "Correlates of Reading Improvement," *Journal of Developmental Reading*, 7 (Autumn 1963), 29-39.
19. Lord, Frederic M. "Elementary Models for Measuring Change," in Chester Harris (Ed.), *Problems in Measuring Change*. Madison: University of Wisconsin Press, 1963, 21-38.
20. Lotsof, E. J., et al. "A Factor Analysis of the WISC and Rorschach," *Journal of Projective Techniques*, 22 (1958), 297-301.
21. Marksheffel, Ned D. "Therapy: An Interdisciplinary Approach," in J. Allen Figurel (Ed.), *Reading and Inquiry*, Proceedings of the International Reading Association, 10, 1965. Newark, Delaware: International Reading Association, 197-200.
22. McLean, Terry Keith. "A Comparison of the Subtest Performance of Two Groups of Retarded Readers with Like Groups of Nonretarded Readers on the Wechsler Intelligence Scale for Children," *Dissertation Abstracts*, 24 (1963), 4800-4801.
23. Monroe, M. *Children Who Cannot Read*. Chicago: University of Chicago Press, 1932.
24. Neville, Donald. "A Comparison of the WISC Patterns of Male Retarded and Nonretarded Readers," *Journal of Educational Research*, 54 (1961), 195-197.
25. Rabinovitch, Ralph D. "Reading and Learning Disabilities," *American Handbook of Psychiatry*. New York: Basic Books, 1959, 857-869.
26. Reid, William Resa. "Psychological Subtest Patterns and Reading Achievement," *Dissertation Abstracts*, 24 (1963), 2366-2367.
27. Sheldon, William D., and Lawrence Carrillo. "Relationship of Parents, Home, and Certain Development Characteristics to Children's Reading," *Elementary School Journal*, 52 (1952), 262-270.
28. Sinks, Naomi, and Marvin Powell. "Sex and Intelligence as Factors in Achievement in Reading in Grades 4-8," *Journal of Genetic Psychology*, 106 (1965), 67-79.
29. Spache, George D. *Diagnostic Reading Scales Manual*. California Test Bureau, 1963a, 5-15.
30. Spache, George D. *Toward Better Reading*. Garrard, 1963b.
31. Strang, Ruth. "Relationships between Certain Aspects of Intelligence and Certain Aspects of Reading," *Educational and Psychological Measurement*, 3 (1943), 355-359.
32. Stroud, J. B., P. Blommers, and Margaret Lauber. "Correlation of WISC and Achievement Tests," *Journal of Educational Psychology*, 48 (1957), 18-26.
33. Stroud, J. B., and E. P. Lindquist. "Sex Differences in Achievement in the Elementary and Secondary School," *Journal of Educational Psychology*, 33 (1942), 657-667.
34. Warner, W. Lloyd. *Social Class in America*. New York: Harper and Brothers, 1960, 140-141.

35. Wechsler, David. *Wechsler Intelligence Scale for Children Manual.* New York: Psychological Corporation, 1949, 2-6.
36. Wilson, A. B. "The Effect of Residential Segregation upon Educational Achievement and Aspirations," unpublished doctoral dissertation, University of California at Berkeley, 1960.
37. Wilson, Robert M. *Diagnostic and Remedial Reading.* Charles E. Merrill Books, 1967.

A Longitudinal Study of Constancy of Reading Performance

KENNETH D. HOPKINS
University of Colorado
AND
GLENN H. BRACHT
Southern Illinois University

Part 1: The Problem, Method, and Conclusions

IS EARLY SUCCESS IN READING related to long-term reading competency? Are differences in initial reading success the result of age and maturational differences that dissipate with time? Are "slow starters" just immature pupils who eventually achieve normally or do they continue to have "ignition" trouble? Although there has been considerable research on the question of IQ constancy, it is surprising that so little attention has been devoted to the related issue of achievement stability, a topic of much greater educational and social importance.

Bloom's survey of research on the predictability of achievement data failed to locate any research that studied even short-term consequences of grade on reading performance. The published studies on the topic of achievement constancy in the area of reading revealed by the literature search are summarized in Table 1. The information in Table 1 reveals that most studies have utilized small Ns (which yield unreliable stability estimates) and that all published studies have extended over a five-year interval or less.

The purpose of this study was to investigate the stability of reading performance as measured by standardized tests at various intervals over the initial eleven grade levels.

Part 2: Findings

Summary

The stability of reading vocabulary and comprehension was studied over the grade one to grade eleven interval using three large

TABLE 1

STABILITY OF READING PERFORMANCE STUDIES

Test	*N*	*Grades*	r	
Metropolitan Reading	105	2-5	.76	Townsend (1944)
Stanford Reading	47	2.9-6.9	.67	Hildredth (1936)
Stanford Paragraph Meaning	81	5-6	.77	
ITBS Reading	27	6.9-8.9	.75	Kvaraceus and Lanigan (1948)
Cooperative Reading	36	7-12	.77	Traxler (1950)
Comprehension	36	8-12	.76	
	36	9-12	.82	
	36	10-12	.82	
Nelson Denny Reading	517	13-14	.83	Silvey (1951)
ITBS and ITED	256	7-9	.83	Krantz (1947)
	251	7-11	.79	
ITBS Reading	900	3-8	.77	
ITBS Vocabulary			.76	
ITBS Reading	9972	5-8	.79	
ITBS Vocabulary			.83	

samples of students. Substantial long-term stability was reflected in both types of tests; grade one scores correlated above .5 with all subsequent measures. By the end of the primary grades, students' scores correlated above .70 with all subsequent measures. When the coefficients were correlated for attenuation to allow an estimate of the relationships after errors of measurements on the test were removed, the values were higher by about .10.

Early performance in reading does not represent temporary maturational status for most pupils but has substantial relationship with terminal achievement levels in both reading vocabulary and comprehension.

Reading tests from three popular achievement batteries were used in the study: Metropolitan Achievement Tests (MAT), Iowa Tests of Basic Skills (ITBS), and Iowa Tests of Educational Development (ITED). The use of different achievement tests is both a strength and a weakness. Varying the tests increases the generalizability of

findings—showing that the results are not limited to the particular measuring instrument in question. At the same time the nature of the variables being measured differs somewhat among the tests: hence the stability estimates will be conservative in nature.

Tests were administered annually in grades one through seven, and also in grades nine and eleven. A reading vocabulary test is included in all three test batteries at each grade level. In the MAT battery the test is called Word Knowledge. A reading comprehension test is included in all three test batteries. However, three reading comprehension tests are included in the ITED battery: Ability to Interpret Reading Materials in the Social Studies, Ability to Interpret Reading Materials in the Natural Sciences, and Ability to Interpret Literary Materials; hence the standardized scores on the three tests were pooled to obtain a composite reading score for each student.

Three independent samples of students were included in the study. The means and standard deviations of the reading vocabulary and comprehension scores are reported for each grade level in Table 2, along with sample sizes and dates of testing. The standard deviations and means are based on the scores of all students in a grade level (cf. Table 1 for sample sizes). In order to determine whether the degree of variability in the samples differed from the population variability, the standard deviations of grade-equivalent scores were compared with corresponding values for the norming population reported in the test manuals. (MAT and ITED estimates were computed from S_o and r_{11}: $S = S_o/\sqrt{1 - r_{11}}$.) A comparison of these values with the sample values revealed the sample values showed greater variability in eighteen of thirty-nine instances and less variability in the remaining twenty-one comparisons. In most cases the differences were small.

The standard deviations were also computed for only the students present at the most extreme grade levels to assess potential selection effects on variability within the samples. The average variability of this smaller sample differed only slightly (about $.03\sigma$ in grade one and $.02\sigma$ in grade eleven) from the total sample within each grade level; hence the stability coefficients are not nonrepresentative as a consquence of atypical sample variability.

TABLE 2

GRADE LEVELS, SAMPLE SIZES, MEANS AND STANDARD DEVIATIONS, AND DATE OF TESTING FOR SAMPLES I, II, AND III

		Vocabulary[a]			*Comprehension*[b]			
Sample	*Grade*	N	$\overline{X}$[c]	*S*	$\overline{X}$	*S*	*Date of Testing*	*Test and Form*
I.	3	461	4.64	1.37	4.79	1.49	3/60	MAT, Elem.
	4	452	4.92	1.31	5.08	1.66	9/60	ITBS, 1
	5	71	6.36	1.57	6.32	1.79	9/61	ITBS, 2
	6	1024	7.37	1.73	7.31	1.53	9/62	ITBS, 1
	7	1065	8.56	1.62	8.39	1.58	10/63	ITBS, 2
	9	1116	16.36	5.52	15.33	5.30	11/65	ITED, X4
	11	919	19.83	5.42	18.77	5.83	12/67	ITED, Y4
II.	1	540	2.10	.50	2.12	.62	3/60	MAT, Prim. I, A
	2	520	3.35	.91	3.63	.85	3/61	MAT, Prim. II, B
	3	1115	4.65	1.27	4.72	1.38	3/62	MAT, Elem. B
	4	1115	4.92	1.14	5.07	1.52	9/62	ITBS, 1
	5	1185	6.30	1.51	6.28	1.76	10/63	ITBS, 2
	6	1195	6.91	1.48	6.90	1.47	10/64	ITBS, 3
	7	1240	8.14	1.57	7.81	1.56	12/65	ITBS, 4
	9	1050	16.06	5.22	15.30	5.30	12/67	ITED, X4
III.	1	1320	2.18	.53	2.19	.65	3/63	MAT, Prim. I, A
	2	1250	3.45	.92	3.63	.90	3/64	MAT, Prim. II, B
	3	1275	4.80	1.25	4.88	1.43	3/65	MAT, Elem. A
	4	1315	4.77	1.11	4.71	1.28	10/65	ITBS, 1
	5	1140	5.84	1.25	5.97	1.38	9/66	ITBS, 2
	6	1095	6.92	1.44	6.98	1.47	9/67	ITBS, 3

[a] "Word Knowledge" test on the MAT.

[b] "Reading" tests of the MAT and ITBS; average of three reading comprehension tests of the ITED: Ability.

[c] Grade equivalent units on the MAT and ITBS; standard scores on the ITED.

Achievement stability

The stability and generalizability coefficients for the reading vocabulary scores are reported in Table 3. Any student who was enrolled in any two or more of the grade levels during the specified years was included in the sample. The N on which the stability coefficient is based is given below the diagonal. For example, the correlation between grade one and grade two reading vocabulary was .82 and was based on 388 pairs of scores. A factor which must be kept in mind in the interpretation of the stability coefficients is the change in test battery at grades four and nine. The slight but generally consistent decreases in stability at grade four are probably related to the change in test battery rather than a real change in the stability of reading vocabulary. The decreases are slight, however, indicating that the stability is not limited to intrabattery inferences but is quite general across competing achievement batteries.

Reading vocabulary scores, which were only moderately stable from grade one through grade three ($r = .6$), suggested considerable fluctuation in pupils' scores on the vocabulary tests. The individual differences in pupils' reading vocabulary did, however, represent a definite lasting characteristic for the group. The correlation of .51 between vocabulary scores at grade one and grade nine indicates that, on the average, a pupil tended to be only half as far from the grade nine mean as he was from the grade one mean. Notice that the stability coefficients tend to increase as grade level is increased. Notice also that there is little change in stability after two or three years. The stability of reading vocabulary scores in grade two was considerably greater than that reflected in grade one performance, correlating about .6 with scores seven years later. Beginning at grade three, the stability of reading vocabulary achievement was maintained at a high level through the secondary grades with an eight-year stability of .76 in Sample I. The stability of reading vocabulary achievement was extremely high for all groups by the beginning of grade five, with the stability coefficients approaching the reliabilities of the test. The generally higher stability coefficients in Sample I can be explained partially by the slightly greater variability of scores (cf. Table 2).

TABLE 3

STABILITY COEFFICIENTS (ABOVE DIAGONALS) AND CORRESPONDING Ns (BELOW DIAGONALS) FOR READING VOCABULARY SCORES FOR SAMPLES I, II, AND III

		Tests and Grade Level									
		MAT			ITBS				ITED		
Grade	*Test*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*9*	*11*	r_{tt}[a]
Sample I											
3	MAT[a]				.82	.86	.81	.79	.79	.76	.95
4	ITBS			388		.80	.80	.76[b]	.78	.77	.91
5	ITBS			368	394		.89	.87	.85	.83	.91
6	ITBS			352	373	776		.88	.87	.85	.91
7	ITBS			341	362	707	891		.88	.87	.89
9	ITED			314	324	611	751	836		.91	.93
11	ITED			285	281	532	642	697	878		.95
Sample II											
1	MAT[a]		.64	.56	.56	.52	.55	.56	.51		.79
2	MAT[a]	415		.78	.72	.65	.66	.66	.59		.92
3	MAT[a]	400	435		.80	.79	.76	.74	.67		.95
4	ITBS	375	400	975[c]		.82	.79	.76[b]	.71		.88
5	ITBS	355	375	885	990		.85	.83	.81		.91
6	ITBS	345	360	825	910	1030		.85	.81		.88
7	ITBS	320	335	765	840	930	1040		.83		.88
9	ITBS	300	305	685	740	795	870	1005			.93
Sample III											
1	MAT[a]		.70	.64	.55	.58	.56				.82
2	MAT[a]	1000		.80	.70	.69	.65				.93
3	MAT[a]	885	995		.78	.78	.76				.95
4	ITBS	815	895	1090		.80	.76				.87
5	ITBS	730	790	940	1070		.83				.86
6	ITBS	720	770	915	1025	1025					.88

[a]From test manuals.

[b]Corresponding value from Merenda and Jackson (1969) was also .76.

[c]There was an increase in sample size at this grade level due to school district reorganization.

Since the correlation coefficients in Table 3 reflect true change and stability of reading vocabulary *plus* errors of measurement, the coefficients were corrected for attenuation to provide an estimate of the stability of true scores in reading vocabulary. These disattenuated stability coefficients are given in Table 4.

TABLE 4

DISATTENUATED STABILITY COEFFICIENTS FOR READING VOCABULARY FOR SAMPLES I, II, AND III

		Grades								
Grade	*Test*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*9*	*11*
Sample I										
3	MAT[a]				.88	.92	.87	.84	.83	.80
4	ITBS					.88	.88	.84	.85	.83
5	ITBS						.98	.97	.92	.89
6	ITBS							.98	.95	.91
7	ITBS								.97	.95
9	ITED									.97
Sample II										
1	MAT		.75	.65	.67	.61	.66	.67	.60	
2	MAT			.83	.80	.71	.73	.73	.63	
3	MAT				.88	.85	.83	.81	.71	
4	ITBS					.92	.90	.86	.78	
5	ITBS						.95	.93	.88	
6	ITBS							.97	.89	
7	ITBS								.91	
Sample III										
1	MAT		.80	.73	.65	.69	.66			
2	MAT			.85	.78	.77	.72			
3	MAT				.86	.86	.83			
4	ITBS					.93	.87			
5	ITBS						.95			

The values in Table 4 provide estimates of the degree of relationship between reading vocabulary performance free from the contaminating effects from errors of measurement and, hence, address the theoretical issue of *true* stability better than the corresponding value found in Table 3. If the reliability of the MAT and ITBS reading

vocabulary tests were able to be increased to 1.0, the correlation between grade one scores would be expected to be .71 with grade six scores and .65 with grade nine scores, reflecting substantial long-range implications of initial reading success. True reading vocabulary scores near the end of the primary cycle (@2.7) were very highly related (r's = .83 — .88) to true scores in grade six and grade eleven (r = .80). the rank order of pupil's *true* reading vocabulary changes little after four years of formal reading instruction (i.e., after 5.1), with disattenuated correlations with all measures thereafter approaching .9 or higher.

It can be concluded that reading vocabulary near the end of grade one gives a good indication of the reading vocabulary of pupils ten years later; the indication is excellent after the completion of grade four.

Reading comprehension

The stability coefficients for reading comprehension tests are given in Table 5. Since three different standardized tests were employed, each one operationally defining reading somewhat differently, the coefficients must be viewed as conservative values. They are, however, generalizability coefficients (*1*) which have allowed both time and test battery to vary. The stability coefficients for reading comprehension are very similar to corresponding values for vocabulary given in Table 3, with the mean and mode vocabulary stability coefficients being .02 and .01 larger than corresponding comprehension values.

The disattenuated stability coefficients are given below the diagonal for each sample in Table 5. These values, averaging about .05 less than corresponding values for vocabulary, indicate that there is less true-score stability in comprehension than in vocabulary, although both are very stable after grade three. The stability coefficients agree very closely with those from Linn (*5*) and Merenda and Jackson (*6*) who studied the grade 4-7 and 5-8 intervals, respectively, using the ITBS.

TABLE 5

STABILITY COEFFICIENTS OF READING COMPREHENSION[a] BEFORE (ABOVE DIAGONALS) AND AFTER (BELOW DIAGONALS) CORRECTION FOR ATTENUATION SCORES FOR SAMPLES I, II, AND III

		Tests and Grade Level									
		MAT			ITBS				ITED		
Grade	*Test*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*9*	*11*	r_{tt}
Sample I											
3	MAT				.82	.79	.80	.77	.76	.72	.96
4	ITBS			.86		.79	.78	.78	.77	.73	.96
5	ITBS			.82	.82		.85	.82	.79	.78	.96
6	ITBS			.83	.81	.89		.87	.85	.83	.91
7	ITBS			.82	.83	.85	.91		.85	.85	.92
9	ITED			.80	.82	.84	.90	.92		.87	.93
11	ITED			.76	.77	.82	.87	.91	.94		.95
Sample II											
1	MAT		.59	.58	.59	.57	.50	.53	.53		.81
2	MAT	.68		.71	.66	.65	.65	.62	.58		.90
3	MAT	.66	.76		.77	.76	.75	.70	.66		.96
4	ITBS	.67	.71	.81		.83	.79	.74[b]	.71		.95
5	ITBS	.65	.70	.79	.87		.83	.78	.78		.96
6	ITBS	.58	.72	.80	.85	.89		.82	.80		.91
7	ITBS	.62	.68	.75	.79	.82	.90		.79		.92
9	ITED	.61	.63	.70	.76	.82	.87	.85			.95
Sample III											
1	MAT		.64	.62	.61	.61	.59				.83
2	MAT	.73		.72	.69	.71	.65				.92
3	MAT	.69	.77		.76	.77	.73				.96
4	ITBS	.69	.75	.80		.81	.76				.93
5	ITBS	.69	.77	.81	.87		.84				.93
6	ITBS	.67	.71	.78	.83	.91					.91

[a]Actual test titles: "Reading" for MAT and ITBS, and the average of three reading interpretation tests (tests 5-7) on the ITED.

[b]Corresponding value from Merenda and Jackson (1969) was .77.

Summary

The stability of reading vocabulary and comprehension was studied over the grade one to grade eleven interval using three large samples of students. Substantial long-term stability was reflected in both types of tests; grade one scores correlated above .5 with all subsequent measures. By the end of the primary grades, students' scores correlated above .70 with all subsequent measures. When the coefficients were correlated for attenuation to allow an estimate of the relationships after errors of measurement on the test were removed, the values were higher by about .10.

Early performance in reading does not represent temporary maturational status for most pupils but has substantial relationship with terminal achievement levels in both reading vocabulary and comprehension.

REFERENCES

1. Cronbach, L. J., N. Rajaratnam, and G. C. Gleser. "Theory of Generalizability: A Liberalization of Reliability Theory," *British Journal of Statistical Psychology*, 16 (November 1963), 137-163.
2. Hildreth, Gertrude. "Results of Repeated Measurement of Pupil Achievement," *Journal of Educational Psychology*, 21 (1936), 286-296.
3. Krantz, L. L. "The Relationship of Reading Abilities and Basic Skills of the Elementary School to Success in the Interpretation of the Content Materials in the High School," *Journal of Experimental Education*, 26 (1957), 97-114.
4. Kvaraceus, W. C., and M. A. Lanigan. "Pupil Performance on the Iowa Every Pupil Tests of Basic Skills Administered at Half-Year Intervals in the Junior High School," *Educational and Psychological Measurement*, 8 (1948), 93-100.
5. Linn, R. L. "A Note on the Stability of the Iowa Tests of Basic Skills," *Journal of Educational Measurement*, 6 (1969), 29-30.
6. Merenda, P. F., and R. M. Jackson. "Relationship between Fourth Grade and Seventh Grade Performance on the Iowa Tests of Basic Skills," *Journal of Educational Measurement*, 5 (1968), 163-165 [cf. erratum, 6 (1969), 31].
7. Silvey, H. M. "Changes in Test Scores after Two Years in College," *Educational and Psychological Measurements*, 11 (1951), 494-502.
8. Townsend, Agatha. "Some Aspects of Testing in the Primary Grades," *Educational Records Bulletin*, 40 (1944), 51-54.
9. Traxler, A. E. "Reading Growth of Secondary School Pupils During a Five Year Period." Achievement Test Program in Independent Schools and Supplement Studies, ERB *Bulletin*, 54 (1950), 96-107.

A Comparison of Formulas for Measuring Degree of Reading Disability

ALBERT J. HARRIS
City University of New York

IN THE MORE THAN SEVENTY YEARS since the first clinical description of reading disability was published, there has been general agreement that a reading disability is characterized by a significant discrepancy between potential reading skill and attained reading skill. There has been no consensus, however, about what to use as the measure of reading potential or expectancy, what to use as the measure of reading skill, how to express the relation of reading to expectancy, or how much of a difference must exist for there to be a disability.

Some people question the need for a formula that expresses reading disability in numerical terms. When the problem is obvious and very severe, it may seem a waste of time and energy to compute a disability score. Yet there are many situations in which a generally accepted formula for measuring reading disability would be very helpful. Among these, the most important is to make possible accurate surveys of the incidence of reading disability in the general population and in many subpopulations. The importance of getting such data is recognized by the U.S. Office of Education, which is currently inviting proposals for conducting this kind of research. A second function of such a measure would be to improve the selection of individuals for inclusion in diagnostic and remedial programs. A third area of use would be in many kinds of research projects, particularly those in which matched groups of disabled readers are needed and those in which improvement as a result of treatment has to be measured.

At the present time we have several methods or formulas for expressing degree of disability in numerical terms. As we shall see, it can make quite a difference which procedure one employs.

Measurement of Reading Expectancy or Potential

Reading expectancy or potential is, or should be, an informed estimate of the level of reading achievement that would be most in harmony with the relevant facts known about the individual. In individual case studies the results of ability tests are tempered by knowledge about the child's motivation, the amount and quality of his schooling, his familiarity with standard English, and the adequacy of sociocultural background. In practice, however, intelligence test scores are commonly used as the sole basis for estimating how well the individual ought to be able to read. The mental age has been the most frequently used intelligence score, and often it is converted into a mental grade score by subtracting five years—or better, 5.2 years. However, in one of the procedures considered in this paper, the Bond and Tinker formula (*1*), IQ is the score used.

A very common procedure, and one which I have recommended in the past, is simply to subtract the reading age from the mental age or to change the mental age to a mental grade score and compare directly with reading grade. This is quick, simple, and requires minimal arithmetic. It is based on the assumption, however, that ideally there should be a perfect correlation between intelligence and reading. It ignores the well-known information that there are many other human characteristics, including auditory and visual discrimination skills, rote memory, ability to attend and concentrate, and so forth, which also correlate positively with reading and which tend to grow or improve with increasing age.

In attempting to avoid excessive reliance on intelligence scores, three procedures have combined intelligence scores with other measures. Monroe (*7*) used an average of mental age, chronological age, and arithmetic, all expressed in grade scores. Horn (*3*) found that chronological age was an important factor and combined it with mental age in different proportions for different age groups. I have recently proposed the use of an Expectancy Age in which mental age is given twice the weight of chronological age (*2*).

Chronological age is the dimension in which growth and learning take place; and when it is used in computing an expectancy score, it can represent a composite of the many factors besides intelligence

that can influence a child's growth in reading competence. When it is given half the weight of mental age, the result is similar to that of a regression equation based on a correlation of .67 between mental age and reading.

Bond and Tinker (*1*) recommend computing reading expectancy in the following manner: multiply the child's IQ by the number of years he has been in school and add one year. This figure gives an expectancy grade which can be compared with the child's reading grade. Like the use of MA alone, this grade relies entirely on an intelligence score. This procedure also seems to imply that no learning relevant to reading performance takes place before the child enters first grade. In its results, this procedure agrees quite well with those using age scores for children with mental ability close to average, but it produces markedly different results for children who are near the extremes of the intelligence distribution.

Let us consider a hypothetical case of a child who repeated first grade and is in the middle of second grade and who is eight years, six months old, with and IQ of 75 and an MA of six years five months. According to the Bond and Tinker procedure his expentancy grade is 2.9, high second grade. According to his mental age of six years, five months, he would be expected to be reading a low-to-middle first grade level. For mentally slow children, the Bond and Tinker expectancies are much too high, in my opinion.

One of the key problems in any estimate of reading expectancy is the accuracy of the data on which it is based. Most satisfactory is a well-administered individual intelligence scale, such as the Wechsler Intelligence Scale for Children. If the examiner reports that the resulting scores are probably too low because of low motivation, cultural handicap, or other clinically valid reasons, the scores should not be used as if they were accurate. When verbal and nonverbal scores disagree, it seems preferable to use the higher score rather than the average or total. And when there is clear evidence of marked educational disadvantage, it is risky to attempt to estimate potential for future learning on the basis of mental ability tests, particularly with young children.

So far, this discussion has focussed on the conventional scoring of mental tests in terms of MA and IQ. Standard scores can also be

used, as shown by Malmquist in his Swedish researches (*5*, *6*). Thorndike (*9*) favors the use of a regression equation to compute expectancy scores from mental ability scores. Either of these kinds of procedures has statistical advantages over age and grade scales. This fact is especially true at adolescent and adult levels, since growth in both measured intelligence and in reading tends to slow down during adolescence. At the elementary school level, however, formulas which use age or grade scores can work quite well.

Measuring Reading Achievement for Comparison with Expectancy

For the purpose of comparing reading attainment with reading expectancy, the reading score used should represent as accurately as possible the instructional reading level of the individual. Standardized reading tests vary in their adequacy for this purpose. Sipay (*8*), for example, gave three standardized silent reading tests to fourth grade children. Two of the tests agreed closely with each other in average grade score, and one of these did not differ significantly from the results of an informal reading inventory. The third test, however, had a mean which was .91 grades higher than one test and .73 grades higher than the other. Thus, the third test would not seem suitable for comparison with expectancy.

It is better for two reasons to use a composite of oral reading and silent reading scores than silent reading alone. The first is that difficulties in word identification are central in many reading disabilities, and low comprehension scores may be the result of inability to recognize the words. The other is that poor readers tend to get many of their correct answers on multiple-choice tests by guessing, making their scores on such test less dependable than the scores of normal readers. I have for many years recommended giving oral reading equal weight with silent reading when comparing with expectancy. The oral reading test may be a standardized reading test or a well-constructed oral reading inventory.

Comparing Expectancy with Achievement

There are two main methods for comparing expectancy with reading achievement: subtracting the reading score from the expec-

tancy score or dividing the reading score by the expectancy score. Subtraction has been the common practice for many years, despite the well-known fact that a year of difference in grade or age score has decreasing significance as one goes up the scale. To correct for this condition it is possible to require larger differences at upper age and grade levels. For example, in earlier editions of my book, *How to Increase Reading Ability*, I recommended as minimum differences that MA should be at least six months higher than reading age in the primary grades, at least nine months higher in grades four and five, and at least one year higher in grade six. Some school systems set their minimum difference for acceptance into a remedial program at one year in the elementary grades but two years in secondary school.

A major difficulty with this procedure is that a stepwise standard catches some children in the middle. For example, if the minimum is set at six months in grade three and nine months in grade four, a child whose reading is eight months below expectancy would be considered disabled at the end of third grade but no longer disabled when he returned to school the next fall. While a stepwise standard is better than a uniform standard at all grades, the former is not so satisfactory as a ratio method.

In a ratio method of comparison, the reading score is divided by the expectancy score and multipled by one hundred to do away with the decimal point. When the reading and expectancy scores are equal, the ratio is one hundred. When the reading score is the higher, the ratio is above one hundred and indicates overachievement; when the reading score is the lower, the ratio comes out below one hundred and indicates underachievements.

Monroe (7) divided average reading grade by expectancy grade to get a reading index. If a child had a reading grade of 2.0 and an expectancy grade of 4.0, his reading index was 50. Monroe considered that a reading index below 80 indicated a real disability. Johnson and Myklebust recommended using age scores rather than grade scores. They divide reading age to get a learning quotent (4). They consider a learning quotient below 90 to indicate a disability. If we use the same example and add 5.2 years to each grade score to change it to an age score, we divide 7.2 years by 9.2 and get a learning quo-

tient of 78, which is well below 90 and again indicates a reading disability.

The procedure I recommend in the Fifth Edition of *How to Increase Reading Ability* (2) is called a Reading Expectancy Quotient and has features in common with both the Monroe and the Johnson-Myklebust procedures. It resembles the Monroe procedures in basing the expectancy score on both mental age and chronological age. It resembles the Johnson-Myklebust procedure in using age scores rather than grade scores. It resembles both in being a ratio, with reading score divided by expectancy score.

I have computed numerous examples to find the relationship between the reading expectancy quotient and the subtractive difference between expectancy and reading scores. The difference corresponding to an R Exp Q of 90 increases steadily and smoothly from .8 years and the beginning of second grade to 1.0 years at the beginning of fourth grade and 1.5 years at the beginning of eighth grade. This seems to me to be a reasonable place to put the dividing line between disability and nondisability. However, other critical scores could be chosen instead. If one wants a more stringent criterion, lowering the cut-off point from 90 to 87 makes the corresponding difference 1.0 years at the beginning of second grade, and it increases steadily to a difference of 2.0 years at the beginning of eighth grade. Lowering it still more to 85 corresponds to total inability to read at beginning grade two, a difference of 1.4 years at beginning grade three, and a steady increase to a difference of 2.3 years at beginning grade eight. This point seems to me to be too severe. I, therefore, agree with Johnson and Myklebust in setting the dividing line between disability and nondisability at a reading expectancy quotient of 90.

Whether one uses a subtractive procedure or a ratio procedure, someone has to make an arbitrary decision as to where to put the dividing line between disability and nondisability. Hopefully the decision made will be in agreement with the experience of seasoned remedial specialist. It should function to identify those individuals who need more individualized and intensive help than most classroom teachers can provide. It should not exclude children of limited

mental ability who read well below their own expectancy level nor children of superior mental ability whose reading is passable but mediocre. The reading expectancy quotient seems to have these characteristics when 90 is used as the dividing point between disability and nondisability.

It does make a difference what method or formula one uses and what cut-off point one chooses. The different formulas all come to the same conclusion when there is a severe disability or no sign of difficulty, but there are marginal cases in which the various formulas lead to opposite conclusions. Let us consider James Doe, a hypothetical boy who is ten years, zero months old and in the eighth month of the fourth grade. His MA is 9 years, 0 months; his IQ is 90, and his average reading grade is 3.2. His arithmetic grade is 4.2. Since his reading is 1.6 years below his grade placement, all would agree that he is retarded in reading. Does he have a reading disability? Two of the formulas (Monroe, Bond and Tinker) say "yes"; two formulas (subtraction of reading age from mental age and the Johnson-Myklebust procedure) say "no"; and his reading expectancy quotient is 90, right on the cut-off point. Cases like this bring home the importance of choosing the best possible way to express degree of reading disability in numerical terms.

One of the problems not yet discussed is the question of what to do about bright students who are underachievers. If Joe Smith is a fifth grader with mental ability equivalent to that of an eighth grader, but who reads just at fifth grade level, should he be considered to have a reading disability? Assuming that he is ten years old, his reading expectancy age is 12.0 and his reading expectancy quotient is 85, which is within the reading disability range.

It is desirable, however, to have a way of expressing the difference between a bright youngster with average reading ability and a youngster with severely below grade reading. This distinction can be made with an additional measure called the Reading Quotient. The reading quotient is simply reading age divided by chronological age. Since Joe's reading is average for his age and grade, his reading quotient is approximately 100. When a child has a low reading expectancy quotient but his reading quotient is 90 or above, the proper label is "underachiever" rather than "reading disability."

REFERENCES

1. Bond, Guy L., and Miles A. Tinker. *Reading Difficulties: Their Diagnosis and Correction* (2nd ed.). New York: Appleton-Century-Crofts, 1967, 91-95.
2. Harris, Albert J. *How to Increase Reading Ability* (5th ed.). New York: David McKay Company, 1970, 208-216.
3. Horn, Alice. *The Uneven Distribution of the Effects of Specific Factors,* Southern California Education Monographs, No. 12. Los Angeles: University of Southern California Press, 1941.
4. Johnson, Doris J., and Helmer R. Myklebust. *Learning Difficulties: Educational Principles and Practices.* New York: Grune and Stratton, 1967.
5. Malmquist, Eve. *Factors Related to Reading Disabilities in the First Grade of the Elementary School.* Stockholm: Almqvist & Wiksell, 1958.
6. Malmquist, Eve. *Lässvarigheter på grundskolans lågstudium: Experimentella studier.* Forskningsrapporter från Statens Försöksskola i Linköping Linköping, Sweden: Utbildningsförlaget, 1969.
7. Monroe, Marion. *Children Who Cannot Read.* Chicago: University of Chicago Press, 1932.
8. Sipay, Edward R. "A Comparison of Standardized Reading Scores and Functional Reading Levels," *Reading Teacher,* 17 (January 1964), 265-268.
9. Thorndike, Robert L. *The Concepts of Over- and Underachievement.* New York: Teachers College Press, Columbia University, 1963.
10. Ullmann, Charles A. "Prevalence of Reading Disability as a Function of the Measure Used," *Journal of Learning Disabilities,* 2 (November 1969), 556-558.

The Validity of the Instructional Reading Level

WILLIAM R. POWELL
University of Illinois at Urbana-Champaign

THE REAL VALUE of the informal reading inventory (IRI) lies not so much in its identification of the instructional reading level—and, by interpolation, the independent and frustration levels—rather, its real value is that it affords the possibility of evaluating reading behavior in depth. Furthermore, it has the potential for training prospective teachers about reading behavior, a potential unequaled by other types of learning opportunities. For purposes of training teachers, the process becomes the product.

The strength of the IRI is not as a test instrument but as a strategy for studying the behavior of the learner in a reading situation and as a basis for instant diagnosis in the teaching environment.

What we are really concerned with is the degree of mastery. The child does not have an *instructional* level; he has only a *performance* level. To obtain the desired performance level, adjustment has to be made in the criterion levels, the learning time, or the linguistic complexity of the written language. The selection of the adjustment variables is a teacher task and, therefore, an instructional one.

When we speak of instructional level, we are referring to a teacher task; when we speak of performance, we are referring to the learner's behavior; and when we speak of difficulty of material, we are referring to the characteristics of the media. For maximum learning, all three have to match: performance level (child), instructional level (teacher), and passage difficulty (material). The instruction should be provided by the teacher at the performance level of the child that will allow for the exclusion of interfering or disruptive reading behaviors.

Background

Statements and comments on the informal reading inventory are not new. Indeed, many papers on this general topic have been

presented. But in the past few years, the nature of the discussion has shifted from one of description and exposition to one of inquiry and critical analysis. This altered perspective now is focusing on the critical issues—generating critical questions in an open forum about the concept, criteria, application, and empirical basis of the IRI, which has become a part of the fabric of reading instruction since its structured formulation by Betts (*2*) nearly thirty years ago.

A major product derived from the use of the IRI is the identification of three distinct reading levels—independent, instructional, and frustration. For instructional purposes, the assumption has been that each literate individual, regardless of maturity, has three such levels. Supposedly, these would be in hierarchical order in relationship to the difficulty of the materials, with the independent reading level being the lowest, or easiest, of the three. The other two levels, instructional and frustration, follow in ascending order as the readability of the material increases. Each reading level is alleged to have specific instructional implications for the classroom teacher. While the existence of three different reading levels for literate persons is a powerful concept, it would have to be considered presently as a functionally useful but unvalidated construct.

Because the use of an IRI embodies most of the elements of the instructional environment, this process offers potential beyond the important task of making a match between children and suitable materials. There is the opportunity for teachers to gain diagnostic insights, from the simple indication of level to the complex evaluation of reading behavior. The latent power of this process is just beginning to be tapped as a means of expanding the conceptual framework of individuals in teacher education programs.

Purpose and Limitations

Contrary to the possible implications of the title of the paper, I shall explore some of the facets and perceptions beyond the limited range of the instructional level. I trust the fact that I do not expand broadly into other related dimensions of the IRI will not be taken as a lack of sensitivity to the probable issues there. Components such as

comprehension, rate, and symptoms of difficulty all play their interacting parts in affecting the total reading performance.

Rather than elaborate on the descriptive elements of the informal reading inventory, I am going to assume that you are somewhat familiar with its characteristics, construction, and administration, as well as with at least one scoring scheme used for the interpretation of levels. For those who wish to pursue information about the fundamental constituents of the IRI, I would refer you to Betts (*2*), Johnson and Kress (*10*), and Zintz (*18*).

The purpose of this paper is to present a critical inquiry about the product of the informal reading inventory and about some of the elements used in the process of determining that product. To achieve this purpose I propose to review recent developments on this topic briefly and to raise three particular questions. The first two deal with the process of the IRI, and the last deals with its product:

1. What is a suitable criterion level for word recognition in identifying the instructional reading level?
2. Is it appropriate to apply one set of performance standards uniformly across all grade levels?
3. Could it be that the major product of the IRI, i.e., the identification of three distinct reading levels, is a misinterpretation?

Recent Inquiries

Without much doubt, the most widely used predetermined standards for evaluating reading performance on the IRI are those originally suggested by Betts (*2*). His criteria follow:

Level	*Word Recognition (%)*	*Comprehension (%)*	*Symptoms of Difficulty*
Independent	99	90	none
Instructional	95	75	none
Frustration	90	50	some

Through the years, several individuals have expressed reservations and concern about the original criteria, but few have suggested

TABLE 1

REVISED SCORING CRITERIA FOR THE INFORMAL READING INVENTORY (IRI)

Passages 1-2

Word Recognition

Independent	Instructional	Frustration
1/99-1/50	1/49-1/8	1/7 (and below)

Comprehension

100%-90%	89%-70%	69% or less

Passages 3-5

Word Recognition

Independent	Instructional	Frustration
1/99-1/50	1/49-1/13	1/12 (and below)

Comprehension

100%-90%	89%-70%	69% or less

Passages 6+

Word Recognition

Independent	Instructional	Frustration
1/99-1/50	1/49-1/18	1/17 (and below)

Comprehension

100%-90%	89%-70%	69% or less

TABLE 2

WORD RECOGNITION ERROR RATIOS BY EIGHT SETS OF CRITERIA

Criteria Levels	*Powell*	*Spache*	*Durrell*	*Gilmore*	*Gray*	*Gates McKillop*	*Betts-Killgallon*	*Cooper*
P		1/4	1/3	1/3	1/7		1/20	1/50
1[2]	1/6	1/5		1/5	1/8		1/20	1/50
2[1]		1/8	1/8	1/6	1/11		1/20	1/50
2[2]	1/8	1/7	1/9			1/2	1/20	1/50
3[1]		1/10		1/8	1/11		1/20	1/50
3[2]	1/11	1/13	1/12			1/3	1/20	1/50
4	1/13	1/15	1/13	1/11	1/10	1/4	1/20	1/25
5	1/12	1/16	1/16	1/13	1/11	1/6	1/20	1/25
6	1/17	1/16	1/18	1/14	1/9	1/6	1/20	1/25
7		1/16	1/17	1/18	1/10	1/6	1/20	
8		1/18		1/20	1/9	1/6	1/20	

other standards of performance that differed markedly.• In 1968, at the IRA conventon in Boston, I broke the "silence of doubt" and openly challenged the existing sets of criteria (*12*). My investigation suggested that the original criteria simply are not consistent with the actual reading behavior of children. The Betts criteria for the word-recognition dimension in evaluating oral reading behavior for the instructional reading level are too stringent, even for the proficient readers. The alternate criteria that I found to be more consistent with children's actual performances are presented in Table 1.

At the IRA convention in Kansas City, one full symposium program was devoted to the validity of the IRI. These presentations have subsequently been published (*8*). Particularly noteworthy out of that symposium collection was a paper by Beldin (*1*). He systematically traced the historical development of the informal reading inventory and pointed out some of the issues regarding the process of this instrument.

• Smith (*15*) is a notable exception to this statement. Since 1959, she has proposed a lower percentage for correct pronounciation. Smith suggests an 80 to 85 percent accuracy range. Spache (*16*) has also offered an opinion that the Betts standards are arbitrarily too high.

In November 1969 at the NCTE convention in Washington, D. C., Dunkeld and I (*13*) presented further comparative data concerning the validity of the criteria. We compared sets of criteria from eight sources, five of which were derived from commonly used oral reading tests. These data are presented in Table 2. Attention should be called to the similarity of the criteria in the first four columns. Also, please note that only one of the word-recognition error ratios (on the Gilmore at the eighth grade) reached the predetermined standards originally set by Betts.

REFERENCES TO TESTS IN TABLE 2.

DURRELL, DONALD D. *Durrell Analysis of Reading Difficulty*. New York: Harcourt, Brace and World, 1955.

GATES, ARTHUR I., AND MCKILLOP, ANNA S. *Gates-McKillop Reading Diagnostic Tests*. New York: Bureau of Publications, Teachers College, Columbia University, 1962.

GILMORE, JOHN V. *Gilmore Oral Reading Test*. New York: Harcourt, Brace and World, 1968.

GRAY, WILLIAM S. *Gray Oral Reading Tests*, H. M. Robinson (Ed.). Indianapolis: Bobbs-Merrill, 1963.

SPACHE, GEORGE D. *Diagnostic Reading Scales*. Monterey, California: California Test Bureau, 1963.

Questions and Issues

What is a suitable criterion level for word recognition in identifying the instructional reading level? We have enough evidence to suggest what is an unsuitable criterion but not enough yet to say with assurance what is suitable. It definitely would appear that the original criterion of 95 percent correct pronunciation (word recognition)—that is, one error in every twenty running words—is too high for all age-grade levels.

The way two occurrences relate tends to support this conclusion. Studies have been conducted to evaluate other concurrent events using the original criteria, such as investigations comparing grade placement scores derived from standardized reading measures with levels obtained from the informal reading inventory. In general, such studies have consistently indicated that scores from standardized tests

vary at least from one-to-three years *above* a reported instructional reading level, as determined by the IRI.* While one study did clearly caution that generalizing from standardized scores to the instructional reading level was tenuous at best, a significant gap did exist between the two types of assessment for a large number of the children studied (*6*). Undoubtedly, the nature of the assessment process between the two types of instruments could be expected to produce a difference between scores. Nevertheless, the degree of difference has been viewed with some suspicion as being greater than what should be expected for proficient readers.

Now, suppose we apply this information to the model generally used for determining reading disability. The model typically used is the degree of difference between the subject's estimated capacity and actual reading achievement, as determined by scores from a standardized test. If the difference between capacity and achievement equals or exceeds a predetermined cut-off point, then the child is said to be disabled. If we apply the difference between standardized reading achievement measures and the instructional reading level and then add the discrepancy between estimated capacity scores and the reading achievement scores, an interesting phenomenon occurs. For most children of average ability with at least average reading achievement scores, their instructional reading level is not likely to be within the acceptable lower limits of their estimated capacity. Suppose the estimated capacity and reading achievement match perfectly; even then, the difference between their reading performance, as estimated by standardized tests, and their instructional reading level, as measured by the IRI, would be great enough to cause the instructional reading level to be outside the usually acceptable limits of normal reading behavior. Is this outcome at all suitable? If the

* The studies by Killgallon (*11*), Daniels (*5*), Williams (*17*), Sipay (*14*), Davis (*6*), and Brown (*3*) all support the contention that standard tests tend to overestimate the instructional level. All studies except the one by Sipay used the Betts criteria with slight modification. For example, Williams adjusted the minimum acceptance in comprehension at the instructional level from 75 to 70 percent. Sipay, however, used the criteria suggested by Cooper (*4*) [see Table II]. Since these criteria are even more rigorous than those developed by Betts, the same pattern was found.

criteria for determining the instructional reading level were representative of children's actual reading performance, would the discrepancies noted diminish? It would seem logical to assume that for youngsters of average ability and achievement, the instructional level should be within the tolerable limits of their estimated capacity.

Is it appropriate to apply one set of performance standards uniformly across all grade levels? Here, we need to divide our attention between the quantitative and the qualitative. The quantitative dimension refers to the numerical count of the errors or miscues used in computing the percent correct figure or the word recognition error ratio. The qualitative aspect of the issue refers to the types of errors or miscues that are permitted for computational purposes.

The data in Table 2 would not support an assumption that the same quantitative ratio or percentage figure can apply uniformly across all grade levels. Apparently, there is a differential function in oral reading miscues from grade level to grade level.

My earlier investigation, resulting in new criteria, implied that the change in the word recognition error ratio was due to the age/grade of the child. While the maturity of the reader certainly would be a factor in such a shift of error ratio, I now believe that the important factor is not the age/grade relationship but the difficulty level of the passage.

The implications of this conclusion were made only too clear to me by a written comment from one of my graduate students:

> If we now decide to use the criteria for passage levels rather than the child's level in school, is our decision to do so founded on the evidence in your study? For the average child's reading grade, it won't make much difference, but what about the sixth grader referred to the clinic experiencing difficulty in reading. On which basis do we judge his performance, on say first and second grade passage? There is a big difference between 1/8 and 1/18.*

All available data, nevertheless, seem to indicate that there is an inverse relationship between the difficulty, or readability, level of a

* Comment by Patricia Stoll as contained in an intraoffice memo to the author.

passage and the number of word recognition errors tolerated by a reader. That is, the easier the material, the higher the percent of miscues that can be permitted by the reader while still maintaining an acceptable understanding level of the material read. Conversely, the more complex the written language, the fewer the number of deviations that can be so tolerated and still realize an acceptable comprehension level.

The key word in such a discussion as the one in the preceding paragraph is *tolerate*. (*1*) What is meant by *tolerate*? It is the level of error difficulty or deviation from the expected response that is not detrimental to total reading performance. The tolerance level allows for a compensation or adjustment of the reader within *his* range of functioning. As error intolerance increases, the material and instruction must be adjusted downward; and as error intolerance decreases, the adjustment should increase.

Before leaving the quantitative dimension of this issue, I would like to offer a point of curiosity. What relationship exists, if any, between the percent of word recognition deviations and sentence length? As the material increases in complexity and difficulty, the sentence length will also increase. Is there an inverse relationship between sentence length and error tolerance? Or is deep structure or some other linguistic factor the important variable, not sentence length?

Quantitatively, uniformity across passage levels would not appear to be appropriate either. The type of errors that significantly affect a reader's tolerance level are not uniform from level to level. That is to say that the types of significant errors between an average second grader and an average sixth grader are different, and should be. This observation is based on a doctoral study by Dunkeld (*7*) currently near completion at the University of Illinois. It also coincides with the types of findings by Goodman (*9*) in her study of oral reading miscues. She states, "It became evident that the type of miscues which beginning readers made change qualitatively as they become more proficient readers." Therefore, certain types of miscues in the reading of a passage of second grade difficulty might not be scored as errors at that level but might be used for determining error ratios at the fourth grade difficulty level, and vice versa.

An apparent problem concerning the qualitative value placed on errors depends on the definition and classification used in processing those errors. There is little agreement among authorities on what constitutes a substitution, a mispronunciation, etc. The lack of agreement is not only in the basic definition but also in the implications. Certainly, if error types are to have relevance and provide cues for instruction, then a reasonable degree of common interpretation will have to be established.

Could it be that the major product of the IRI, *i.e., the identification of three distinct reading levels, is a misinterpretation?* To search for truth, one has to be willing to risk the ultimate. To critically analyze the process and product of the IRI, one has to consider that the ultimate answer may be negative—that indeed the IRI has no actual validity and that we who work with it are making something out of it that it is not. But that finding would offer positive direction for other types of options.

Research evidence to support the construct of an instructional reading level is minimal and incomplete, as it is for the frustration reading level. This statement does not mean that we do not believe such levels exist. It simply means we do not yet have the data to support our beliefs.

One of the traditional beliefs regarding reading levels is that they form a hierarchical sequence—independent, instructional, and frustration, in that order. Spache challenges that opinion by reversing the position of the instructional and independent reading levels. He orders the levels this way: instructional, independent, and frustration.

There is absolutely no empirical data for defining the rank order nor the limits of the independent reading level. It has been assumed to be beyond the upper limits of the instructional level; therefore, Spache's reversal of the rank order may well be correct. How would we know which sequence is correct?

Since everyone is guessing about the location of the independent reading level, I might as well offer a conjecture on the subject. My impression is that the independent reading level is not static (it "floats"). It may not always be located above or below the instructional reading level. The leverage to the reader is the interest value

of the ideas and concepts. The greater the interest, the higher the passage difficulty can be for the independent reading level of a particular pupil. Conceivably, interest could cause this level to be quite variable, and it may be equal to or above the instructional level in specific types of materials. It is possible that for brief, transitory high-intensity periods, the interest value could project the independent reading level into the usual frustration zone (defined as beyond the lower limits of the instructional level). But until we have some data with which to define the limits of the independent level, your guess is as good as the three just given.

Another option may be that we are applying the right labels to the wrong agent. What we are really concerned with is the degree of mastery. The child does not have an *instructional* level; he only has a *performance* level. To obtain the desired performance level, adjustment has to be made in criterion levels, the learning time, or the linquistic complexity of the written language. The selection of the adjustment variables is a teacher task and, therefore, an instructional one. When we speak of instructional level, we are referring to a teacher task; when we speak of performance, we are referring to the learner's behavior; and when we speak of difficulty of material, we are referring to the characteristics of the media.

For maximum learning, all three have to match: performance level (child), instructional level (teacher); and passage difficulty (material). The instruction should be provided *by the teacher* at the performance level *of the child* that will allow for the exclusion of interfering or disruptive reading behaviors.

Concluding Statement

The value of the IRI lies not in its identification of what has been called the instructional level (and the other levels by interpolation) because there are probably more effective and efficient methods of accomplishing such tasks. The use of cloze procedure is one alternative already available that has a considerable body of research data to support it. The real value of the IRI is that it affords the possibility of *evaluating reading behavior in depth*. Furthermore, it has the potential for training prospective teachers about reading

behavior, a potential unequaled by other types of learning opportunities. For purposes of training teachers, the process becomes the product.

The strength of the IRI is not as a test instrument but as a strategy for studying the behavior of the learner in a reading situation and as a basis for instant diagnosis in the teaching environment.

REFERENCES

1. Beldin, H. O. "Informal Reading Testing: Historical Review and Review of the Research," in William K. Durr (Ed.), *Reading Difficulties: Diagnosis, Correction, and Remediation.* Newark, Delaware: International Reading Association, 1970, 67-84.
2. Betts, Emmett A. *Foundations of Reading Instruction.* New York: American Book Company, 1946.
3. Brown, Sandra R. "A Comparison of Five Widely Used Standardized Reading Test Scores and an Informal Reading Inventory for a Selected Group of Elementary Children," unpublished doctoral dissertation, University of Georgia, 1963.
4. Cooper, J. Louis. "The Effect of Adjustment of Basal Reading Materials on Reading Achievement," unpublished doctoral dissertation, Boston University, 1952.
5. Daniels, John E. "The Effectiveness of Various Procedures in Reading Level Placement," *Elementary English,* 39 (October 1962), 590-600.
6. Davis, Sister M. Catherine Elizabeth. "The Relative Effectiveness of Certain Evaluative Criteria for Determining Reading Levels," unpublished doctoral dissertation, Temple University, 1964.
7. Dunkeld, Colin G. "The Validity of the Informal Reading Inventory for the Designation of Instructional Reading Levels: A Study of the Relationships between Children's Gains in Reading Achievement and the Difficulty of Instructional Materials," unpublished doctoral dissertation, University of Illinois, 1970.
8. Durr, William K. (Ed.). *Reading Difficulties: Diagnosis, Correction, and Remediation.* Newark, Delaware: International Reading Association, 1970.
9. Goodman, Yetta. "Studies of Reading Miscues," from translated remarks made in Symposium II, *Applications of Psycholinguistics to Key Problems in Reading.* Kansas City, International Reading Association Convention, 1969.
10. Johnson, Marjorie Seddon, and Roy A. Kress. *Informal Reading Inventories.* Reading Aids Series. Newark, Delaware: International Reading Association, 1965.
11. Killgallon, Patsy A. "A Study of Relationships among Certain Pupil Ad-

justments in Reading Situations," unpublished doctoral dissertation, Pennsylvania State University, 1942.

12. Powell, William R. "Reappraising the Criteria for Interpreting Informal Inventories," in Dorothy L. DeBoer (Ed.), *Reading Diagnosis and Evaluation,* 1968 Proceedings, Volume 13, Part 4. Newark, Delaware: International Reading Association, 1970, 100-109.
13. Powell, William R., and Colin G. Dunkeld. "Validity of the IRI Reading Levels," *Elementary English* (in press).
14. Sipay, Edward R. "A Comparison of Standardized Reading Achievement Test Scores and Functional Reading Levels," unpublished doctoral dissertation, University of Connecticut, 1961. [See also *Reading Teacher,* 17 (January 1964), 265-268.]
15. Smith, Nila Banton. *Graded Selections for Informal Reading: Diagnosis for Grades 1 through 3.* New York: New York University Press, 1959.
16. Spache, George D. *Reading in the Elementary School.* Boston: Allyn and Bacon, 1964.
17. Williams, Joan. "A Comparison of Standardized Reading Test Scores and Informal Reading Inventory Scores," unpublished doctoral dissertation, Southern Illinois University, 1963.
18. Zintz, Miles V. *The Reading Process: The Teacher and the Learner.* Dubuque, Iowa: Wm. C. Brown Company, 1970.